GenPlex Introduction

<Version 3.0>

Istech has all rights of this manual and this product.
You cannot reprint, copy or distribute this manual and this product, without permission of Istech Corp. in advance.
We will consider all, who have installed and who are using our product,
will agree on this policy.

Istech Inc.

Copyright ⓒ 2008 ISTECH Inc.

Table of Contents

Copyright ⓒ 2008 ISTECH Inc. i
Introduction 8
1. GenPlex Introduction 9
1.1. Summary 9
1.2. Main Function 10
1.2.1. Easy Data Importing 10
1.2.2. Preprocessing 10
1.2.3. DEG (Differentially Expressed Gene) Finding 10
1.2.4. Clustering 11
1.2.5. Classification 11
1.2.6. Pathway Analysis 11
1.2.7. Biological Annotation & Data Mining 12
1.3. Recommended Computer Requirements 13
Preprocessing 14
2. Preprocessing 15
2.1. File 15
2.1.1. Import Affymetrix Gene Chip Data 15
2.1.2. Import One-dye Chip Data 20
2.1.3. Import Two-dye Chip Data 21
2.1.4. Open Analysis 24
2.1.5. Recent Analysis 24
2.1.6. Save Analysis 24
2.1.7. Save Analysis As 24
2.1.8. Close Analysis 24
2.1.9. Analysis Properties 24
2.1.10. Configure 24
2.1.11. Exit 24
2.2. Preprocessing 25
2.2.1. Experimental Information 25
2.2.2. Filtering Error Spot 26
2.2.3. Normalization 31
2.2.4. Set Detection 36

2.2.5. Log Transform 37
2.3. Statistics/Plot 38
2.3.1. Statistics 38
2.3.2. Box Plot 39
2.3.3. Histogram 40
2.3.4. MA Plot 41
2.3.5. QQ Plot 42
2.3.6. Correlation Scatter Plot 42
2.3.7. Correlation Matrix Plot 43
2.4. Analysis Data 44
2.4.1. DEG Finding 44
2.4.2. Clustering 45
2.4.3. Classification 46
2.5. Reference 48
DEG Finding 49
3. DEG Finding 50
3.1. File 50
3.1.1. New Analysis 50
3.1.2. Open Analysis 51
3.1.3. Recent Analysis 51
3.1.4. Save Analysis 51
3.1.5. Save Analysis As… 51
3.1.6. Close Analysis 52
3.1.7. Import Data 52
3.1.8. Analysis Properties 54
3.1.9. Exit 54
3.2. Preprocessing 55
3.2.1. Check & Match Data 55
3.2.2. Filter Missing Data 55
3.2.3. Impute Data 56
3.2.4. Log Transform 56
3.3. DEG Finding 57
3.3.1. Fold Change 57
3.3.2. 2-Class Paired Test 65
3.3.3. 2-Class Unpaired Test 66
3.3.4. Multi-Class Test 68

3.3.5. Combine Results 69
3.3.6. Import Gene List 69
3.3.7. Export to Clustering Module 69
3.3.8. Export to Pathway Analysis Module 69
3.3.9. Save Result(s) As Text 70
3.4. Statistics/Plot 71
3.4.1. Basic Statistics 71
3.4.2. Sample Correlation Matrix 71
3.4.3. Box Plot 72
3.4.4. Correlation Scatter Plot 73
3.4.5. Correlation Matrix Plot 73
3.4.6. Venn Diagram 74
3.4.7. Volcano Plot 74
3.5. Reference 77
Clustering 78
4. Clustering 79
4.1. File 79
4.1.1. New Analysis 79
4.1.2. Open Analysis 80
4.1.3. Recent Analysis 80
4.1.4. Save Analysis 80
4.1.5. Save Analysis As 80
4.1.6. Close Analysis 80
4.1.7. Import Data 81
4.1.8. Analysis Properties 82
4.1.9. Exit 82
4.2. Preprocessing 83
4.2.1. Experimental Information 83
4.2.2. Log Transform 83
4.2.3. Gene Filtering 84
4.2.4. Missing Data Filtering 85
4.2.5. Imputation 85
4.2.6. Column Editing 85
4.3. Clustering 88
4.3.1. Hierarchical Clustering 88
4.3.2. K-means Clustering 92

4.3.3. Self Organizing Map 94
4.4. Validation 101
4.4.1. GDI 101
4.4.2. K-value Prediction 102
4.5. Reference 105
Classification 106
5. Classification 107
5.1. File 107
5.1.1. New Analysis 107
5.1.2. Open Analysis 108
5.1.3. Recent Analysis 108
5.1.4. Save Analysis 108
5.1.5. Save Analysis As 108
5.1.6. Close Analysis 109
5.1.7. Load Training Data File(s) 109
5.1.8. Load Test Data File(s) 110
5.1.9. Transpose Data 110
5.1.10. Analysis Properties 110
5.1.11. Exit 110
5.2. Preprocessing 111
5.2.1. Check & Match Data 111
5.2.2. Filter Missing Data 111
5.2.3. Impute Data 112
5.3. Gene Selection 113
5.3.1. Select Algorithm 113
5.3.2. Set Parameter(s) 113
5.3.3. Run 114
5.3.4. Combine Results 116
5.3.5. Set As Active Gene Selection 117
5.3.6. Export to Clustering 117
5.3.7. Export to Pathway Analysis Module 117
5.3.8. Save Result(s) 118
5.4. Classification 119
5.4.1. Select Distance 119
5.4.2. Select Algorithm 119
5.4.3. Set Parameter(s) 120

5.4.4. Classify Test Data 120
5.5. Error Estimation 122
5.5.1. Select Algorithm 122
5.5.2. Set Parameter(s) 122
5.5.3. Run 122
5.5.4. Whole Computation 124
5.6. View 126
5.6.1. Show Sample 3D View 126
5.6.2. Show Summary View 126
5.7. Reference 127
Pathway Analysis 128
6. Pathway Analysis 129
6.1. File 129
6.1.1. New Analysis 129
6.1.2. Open Analysis 130
6.1.3. Recent Analysis 130
6.1.4. Save Analysis 130
6.1.5. Save Analysis As 130
6.1.6. Close Analysis 131
6.1.7. Import Data 131
6.1.8. Analysis Properties 131
6.1.9. Exit 132
6.2. Pathway List 133
6.2.1. Pathway (Image) 133
6.2.2. Pathway (XML) 134
Algorithms 135
7. DEF Finding Algorithm 136
7.1. Fold Change 136
7.2. Two-sample (unpaired) t-test 137
7.3. Volcano Plot 138
7.4. Analysis of Variance (ANOVA) 139
8. Clustering Algorithm 140
8.1. Hierarchical Clustering (HC) 140
8.2. K-means 144
8.3. Self Organizing Map (SOM) 145
9. Classification Algorithm 146

9.1. Gene Selection 147
9.2. Classifier 148
9.3. Generalization Error Estimation 149

Introduction

1. GenPlex Introduction

1.1. Summary

Microarray (DNA Chip) that is able to monitor the intensity of thousands and millions of gene information at the same time, had become the main tool in biotechnology research field in the 21st Century.

The use of Microarray allows us to verify the significant status on the gene level of the gene inside the cell, and through this significant information, we can understand inclusively the relation between these genes.

However, because of the complicatedness of the out-coming data, Microarray requires recent method of all kinds of algorithm and bio-informatics such as Mathematics, Statistics and Computer Science, etc.

GenPlex is a Microarray analyzing software which offers useful information to scientists in analyzing the data suitably to the users in providing various visualization of the results, and also possible to analyze the experiment data through various statistical algorithm.

1.2. Main Function

1.2.1. Easy Data Importing

It is easy to input data for users, because it recognizes automatically various types of raw data.

- Supporting Format
  - Affymetrix Gene Chip Data(CEL)
  - ABI Chip Data
  - Illumina Chip Data(BeadStudio output)
  - GenePix Result
  - ImaGene Data

1.2.2. Preprocessing

It provides the information of the raw data quality through statistical figure and various plot, also provides the Preprocessing function of the data which will be used for future analysis.

- Box Plot
- Histogram
- MA Plot
- QQ Plot
- Sample Correlation Scatter/Matrix Plot: showing the relationship between replications
- Global centering/scaling, Global/Print-tip Lowess, Quantile Normalization, etc.
- Convenient Gene Expression Matrix (GEM format) generation

1.2.3. DEG (Differentially Expressed Gene) Finding

It is possible to apply various conditions of Fold Change, and it also offers us the various statistical analysis methods to compare 2-class or multi-class. It is possible to compare the out-coming DEG using Venn Diagram, and we can verify visually the difference between Fold Change and the statistic analysis result using Volcano Plot.

- Fold Change: One-dye, Two-dye
- Parametric Test for 2-class comparison: Student T-test, Welch's T-test, Z-test
- Nonparametric Test for 2-class comparison: Mann-Whitney test
- Paired Test: Paired T-test, Wilcoxon signed rank test
- Parametric Multi-class Comparison: One way ANOVA
- Nonparametric Multi-class Comparison: Kruskal-Wallis H-test
- Multiple Test Correction: Bonferroni correction, Holm's procedure, Benjamini-Hochberg FDR
- Volcano Plot: Fold Change vs. Statistical Test
- Venn Diagram: combining results from various methods

- Statistics, Box Plot, Correlation Scatter Plot, Correlation Matrix Plot

1.2.4. Clustering

It provides various clustering methods and visualization and it is also possible to verify statistically the clustering result. In case of K-means, it helps the users to conclude (judge) in predicting the most suitable number of cluster.

- Hierarchical Clustering with useful Linkage methods
- K-means Clustering
- SOM (Self Organizing Map) : U-matrix Topographic Profiling
- Statistical Clustering Validation
- K-Value Prediction for K-means Clustering
- Dendrogram with various graphical options for publication

1.2.5. Classification

It is the analysis method mostly used for diagnosis, prognosis and estimation, providing various statistical methods to find out the marker gene. We can avoid the data over-fitting through Generalization Error Estimation of classification, and enables the analysis estimation more accurate and easier with the whole computation.

- Feature selection: finding marker genes for diagnosis
- Classification: classifying samples into pre-defined classes
- Error Estimation: estimating generalized misclassification error rate
- Whole Computation: all-in-one approach for optimal classification
- Sample PCA: powerful visualization with various graphical options for publication

1.2.6. Pathway Analysis

It is able to analyze the biological mutual relationship of genes from DEG Analysis, Clustering Analysis, Classification Analysis, etc., to biological genes. It researches the genes related to common pathway using the biological pathway information of KEGG (Kyoto Encyclopedia of Genes and Genomes) database, and in mapping the DNA Chip expression results, it is understood in the pathway level, the changes of the expression quantity according to the experiment condition.

- Pathway Search: given gene lists, all related KEGG pathways explored
- Pathway Mapping: mapping genes onto pathways
- Up-/Down-regulation display with heatmap

1.2.7. Biological Annotation & Data Mining

For gene group from the statistical analysis result of DEG Finding, Clustering, etc., we can pull out all kinds of biologic information, like GO Annotation, KEGG Pathway, etc. And also can analyze statistically the biological linkage of each group using Gene Ontology.

- Basic Information: NCBI Gene ID, UniGene ID, Gene Symbol, Gene Title, Chromosome Location
- Protein Information: InterPro, Pfam, Prosite, EC Number, Uniprot
- PANTHER Category: PANTHER Family Name, PANTHER Subfamily Name, PANTHER Function, PANTHER Process
- Gene Ontology: GO Molecular Function, GO Biological Process, GO Cellular Component
- Pathway: KEGG Pathway
- ID Conversion: Public ID

1.3. Recommended Computer Requirements

- Microsoft Windows 2000/XP System
- CPU: Pentium 4, higher than 2.4GHz
- RAM: minimum 1GB

Preprocessing

2. Preprocessing

You can have the image file of Microarray experiment results and it is the process of preprocessing of the raw data issued from image scanning.

2.1. File

First, input raw data using Import Data Menu. Import Data Menu is classified in three kinds and is supporting data format as follows:

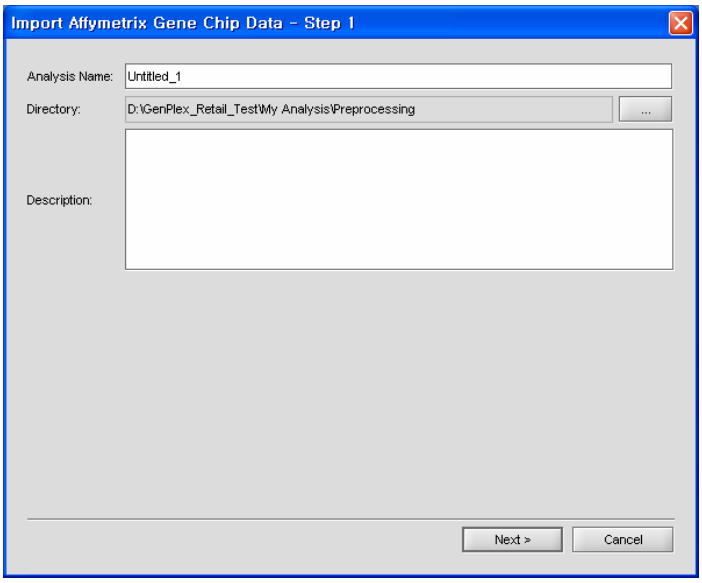
- Affymetrix Gene Chip Data (go to 2.1.1 ▷)
  - CEL File
  - CHP File
- One-Dye Chip Data (go to 2.1.2 ▷)
  - ABI Chip Data
  - Illumina Chip Data(BeadStudio-Exported Gene/Probe Profile Data)
  - Agilent Chip Data(GenePix Results format(*.gpr))
- Two-Dye Chip Data (go to 2.1.3 ▷)
  - GenePix Result
  - ImaGene Data

2.1.1. Import Affymetrix Gene Chip Data

On the menu bar, click [File] → [Import Affymetrix Gene Chip Data], or click the first icon , then Analysis information input window appears.

GenPlex is able to analyze two kinds of data which are 3' IVT Expression Chip and Gene ST array. Among preprocessing procedure of 3' Expression array and Gene ST array, Step 1 and Step 2 are identical but only Step 3 is different from other two Steps.

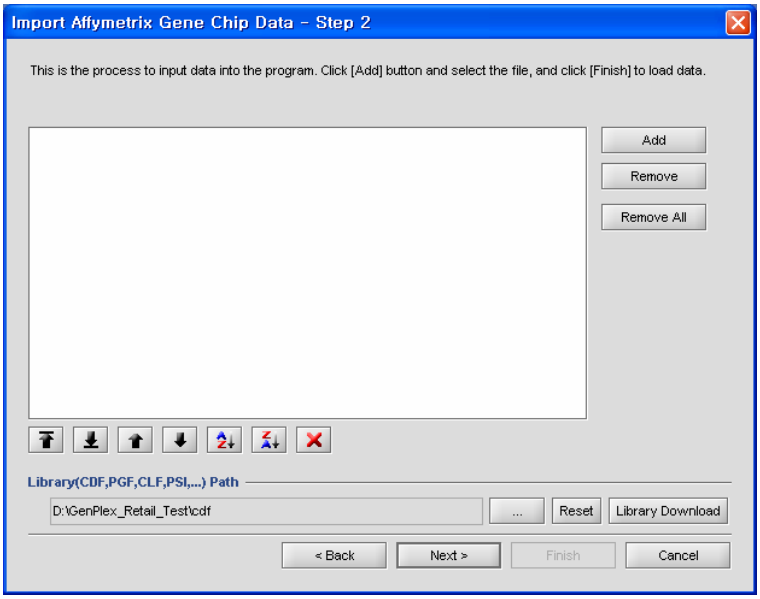
① Step 1



<Figure 2-1> Step 1: Analysis information input window

- Analysis Name: Input the file name of creating Analysis file.
- Directory: Click [...] button to select the location of a new creating Analysis file.
- Description: Input the information on the Analysis (can be omitted).
- Click [Next>] button, then the Analysis will be created, and the data selecting window which is Step 2 will appear.

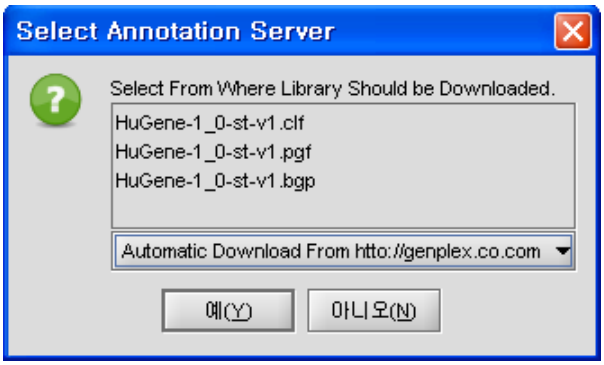
② Step 2



<Figure 2-2> Step 2: Data Selecting Window

- Click [Add] button to select the file to be input, then it will be added to the list on the left side of the window. Click [Remove] and [Remove All] button to delete the item.
- Use buttons to range the files in order.
- Library Path : Select the pathway where the files are saved, .cdf file for Chip type in case of 3' IVT and .clf, .bgp, .pgf file of Chip type in case of ST array. If there is no library file of corresponding Chip type, click [Library Download] button to download the file.
- Click [Next] button, then the preprocessing method window will appear.

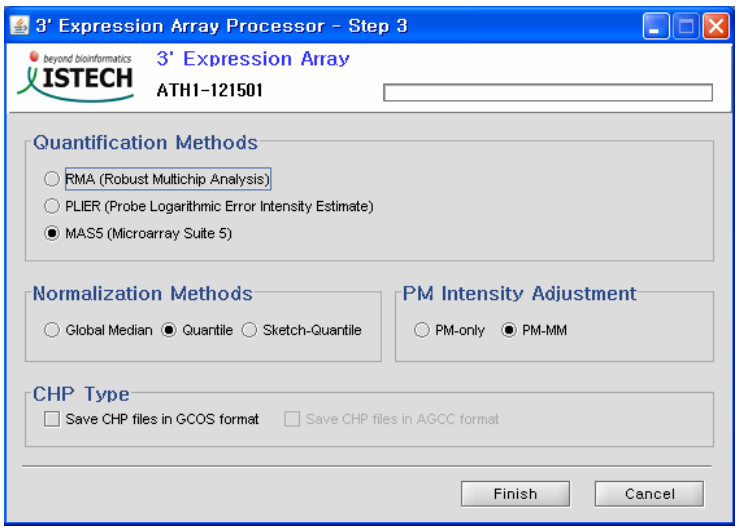
■ Library Download Window: As seen on the <Figure 2-3>, the file list to recognize corresponding Chip Type will be showed, and there are 3 ways to download.
- Automatic Download from http://affymetrix.com
- Automatic Download from http://genplex.co.kr
- Manual Download from http://affymetrix.com



<Figure 2-3> Library Download Window

③ Step 3

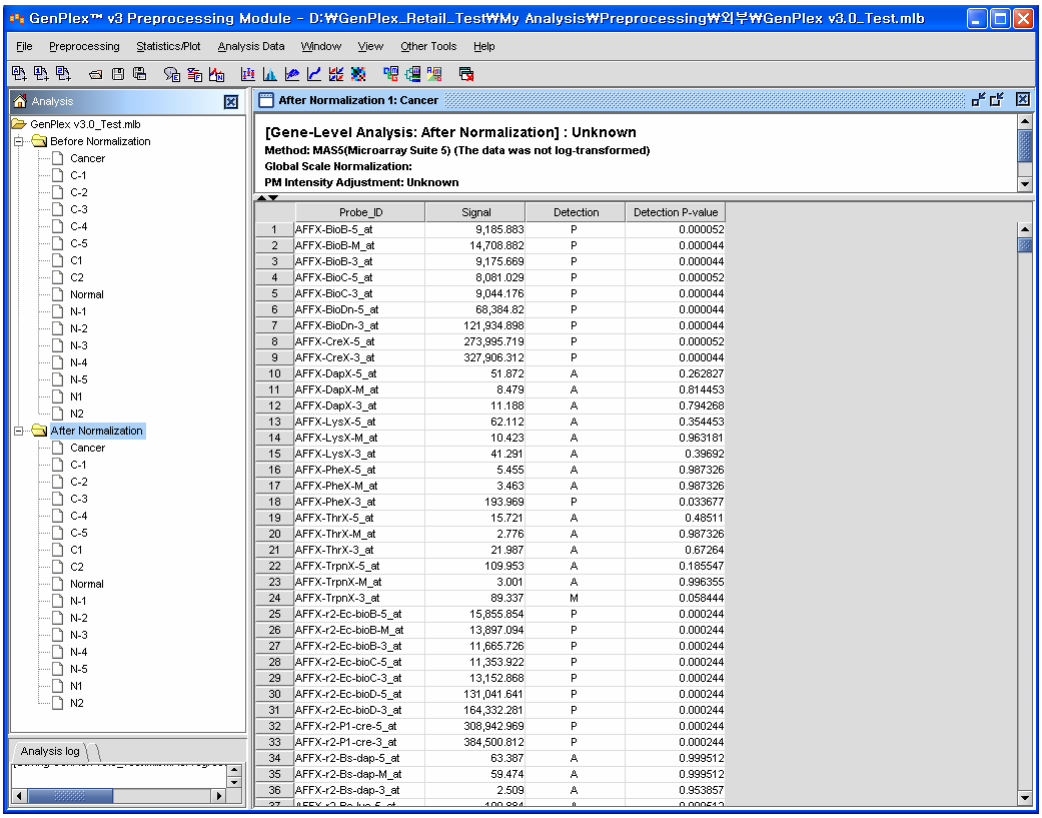
■ 3' IVT Expression Chip



<Figure 2-4> Step 3(3' IVT Expression Chip): Preprocessing Method Selecting Window

- Quantification Methods
  - RMA(Robust Multichip Analysis)
  - Plier[1](Probe Logarithmic Intensity Estimate)
  - MAS5(Microarray Suite 5)
- Normalization Methods
  - Global Median
  - Quantile
  - Sketch-Quantile
- PM Intensity Adjustment
  - PM-only
  - PM-MM
- CHP Type: Select the item of 'Save CHP files in GCOS format' new file .chp will be created in the folder where .cel file is saved.
- Click [Finish] button to operate preprocessing (go to 2.2 ▷).

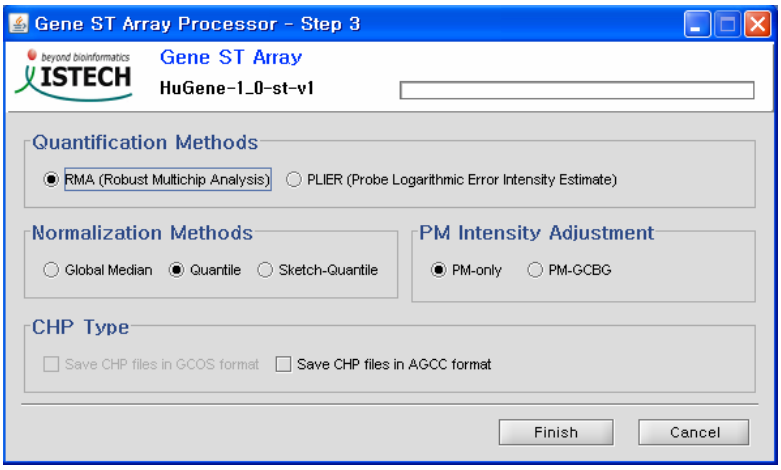
■ 3' IVT Expression Chip Preprocessing Result


GenPlex™ v3 Preprocessing Module – D:₩GenPlex_Retail_Test₩My Analysis₩Preprocessing₩외부₩GenPlex v3.0_Test.mlb
File Preprocessing Statistics/Plot Analysis Data Window View Other Tools Help
Analysis
GenPlex v3.0_Test.mlb
Before Normalization
Cancer C-1 C-2 C-3 C-4 C-5 C1 C2 Normal N-1 N-2 N-3 N-4 N-5 N1 N2
After Normalization
Cancer C-1 C-2 C-3 C-4 C-5 C1 C2 Normal N-1 N-2 N-3 N-4 N-5 N1 N2
Analysis log
After Normalization 1: Cancer
[Gene-Level Analysis: After Normalization] : Unknown
Method: MAS5(Microarray Suite 5) (The data was not log-transformed)
Global Scale Normalization:
PM Intensity Adjustment: Unknown
Probe_ID Signal Detection Detection P-value
1 AFFX-BioB-5_at 9,185.883 P 0.000052
2 AFFX-BioB-M_at 14,708.882 P 0.000044
3 AFFX-BioB-3_at 9,175.669 P 0.000044
4 AFFX-BioC-5_at 8,081.029 P 0.000052
5 AFFX-BioC-3_at 9,044.176 P 0.000044
6 AFFX-BioDn-5_at 68,384.82 P 0.000044
7 AFFX-BioDn-3_at 121,934.898 P 0.000044
8 AFFX-CreX-5_at 273,995.719 P 0.000052
9 AFFX-CreX-3_at 327,906.312 P 0.000044
10 AFFX-DapX-5_at 51.872 A 0.262827
11 AFFX-DapX-M_at 8.479 A 0.814453
12 AFFX-DapX-3_at 11.188 A 0.794268
13 AFFX-LysX-5_at 62.112 A 0.354453
14 AFFX-LysX-M_at 10.423 A 0.963181
15 AFFX-LysX-3_at 41.291 A 0.39692
16 AFFX-PheX-5_at 5.455 A 0.987326
17 AFFX-PheX-M_at 3.463 A 0.987326
18 AFFX-PheX-3_at 193.969 P 0.033677
19 AFFX-ThrX-5_at 15.721 A 0.48511
20 AFFX-ThrX-M_at 2.776 A 0.987326
21 AFFX-ThrX-3_at 21.987 A 0.67264
22 AFFX-TrpnX-5_at 109.953 A 0.185547
23 AFFX-TrpnX-M_at 3.001 A 0.996355
24 AFFX-TrpnX-3_at 89.337 M 0.058444
25 AFFX-r2-Ec-bioB-5_at 15,855.854 P 0.000244
26 AFFX-r2-Ec-bioB-M_at 13,897.094 P 0.000244
27 AFFX-r2-Ec-bioB-3_at 11,665.726 P 0.000244
28 AFFX-r2-Ec-bioC-5_at 11,353.922 P 0.000244
29 AFFX-r2-Ec-bioC-3_at 13,152.868 P 0.000244
30 AFFX-r2-Ec-bioD-5_at 131,041.641 P 0.000244
31 AFFX-r2-Ec-bioD-3_at 164,332.281 P 0.000244
32 AFFX-r2-P1-cre-5_at 308,942.969 P 0.000244
33 AFFX-r2-P1-cre-3_at 384,500.812 P 0.000244
34 AFFX-r2-Bs-dap-5_at 63.387 A 0.999512
35 AFFX-r2-Bs-dap-M_at 59.474 A 0.999512
36 AFFX-r2-Bs-dap-3_at 2.509 A 0.953857

<Figure 2-5> 3' IVT Expression Chip Preprocessing Result Window

[1] Use Affymetrix Power Tools(APT) for RMA, Plier and other calculations
http://www.affymetrix.com/support/developer/powertools/index.affx

■ Gene ST array

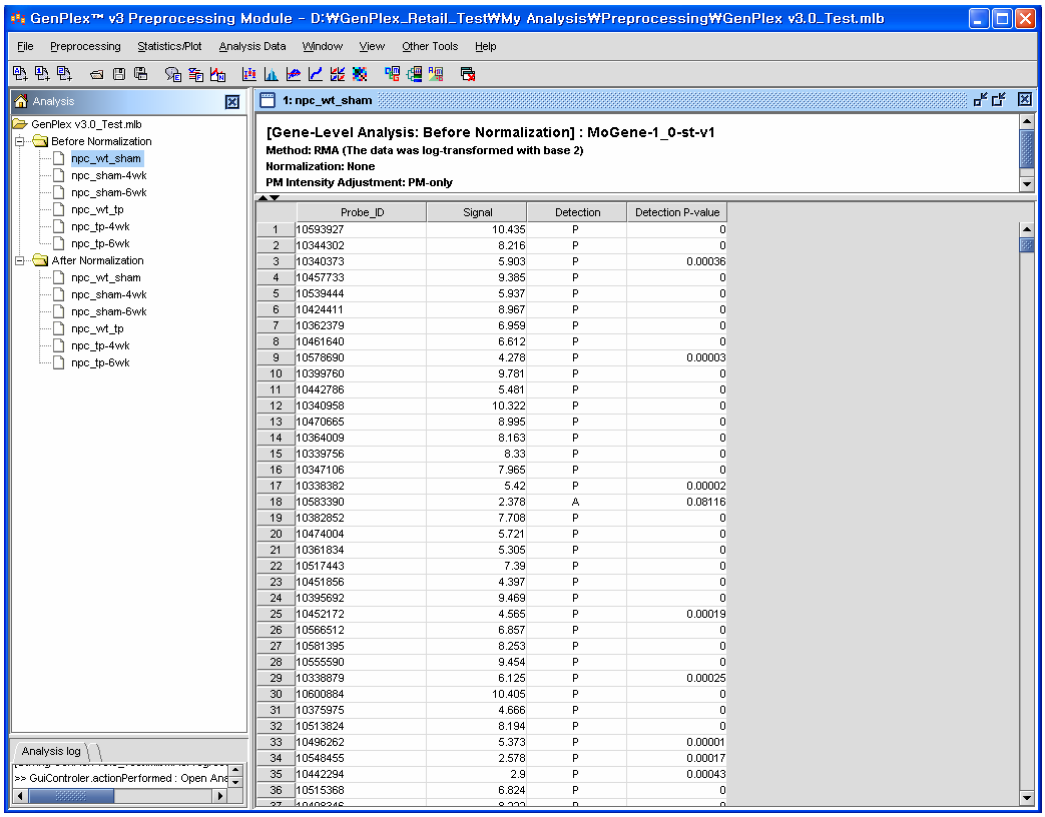


<Figure 2-6> Step 3 (Gene ST array): Preprocessing Method Selecting Window

- Quantification Methods
  - RMA(Robust Multichip Analysis)
  - Plier[2](Probe Logarithmic Intensity Estimate)
- Normalization Methods
  - Global Median
  - Quantile
  - Sketch-Quantile
- PM Intensity Adjustment
  - PM-only
  - PM-GCBG
- CHP Type: Select the item of 'Save CHP files in AGCC format' .chp file will be created in the folder where .cel file is saved.
- Click [Finish] button to operate preprocessing (go to 2.2 ▷).

[2] Use Affymetrix Power Tools(APT) for RMA, Plier and other calculations
http://www.affymetrix.com/support/developer/powertools/index.affx

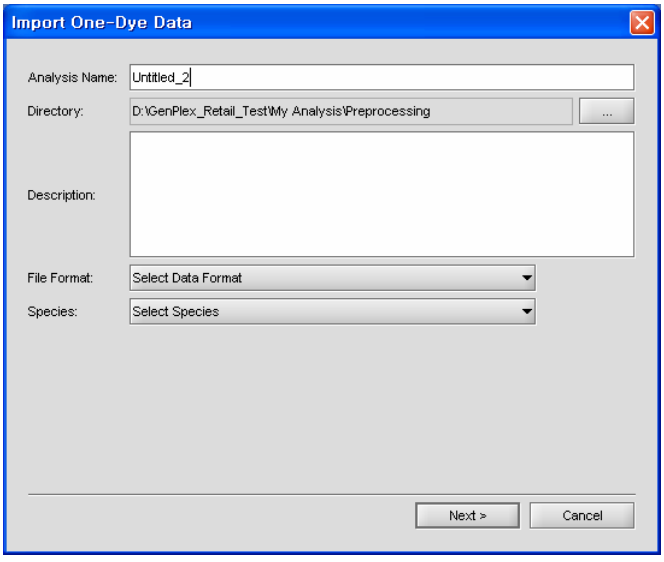
■ Gene ST array Preprocessing Result



<Figure 2-7> Gene ST array Preprocessing Result Window

2.1.2. Import One-dye Chip Data

On the menu bar, select [File] → [Import One-Dye Chip Data], or select the second icon , then Analysis Input Window appears.



<Figure 2-8> One-Dye Chip Data: Analysis Information Input Window

① Analysis Name: Input the name of Analysis File to be created.

② Directory: Click […] button to select the location where Analysis File will be created.

③ Description: Input information on Analysis File (can be skipped).

④ File Format: Select file format of data input.

- ABI chip data
- Illumina chip data(BeadStudio – Exported Gene/Probe Profile Data)
- Agilent chip data(GenePix Results format(*.gpr))

⑤ Species: Select type of the input data (Provides different Species according to Chip kind)

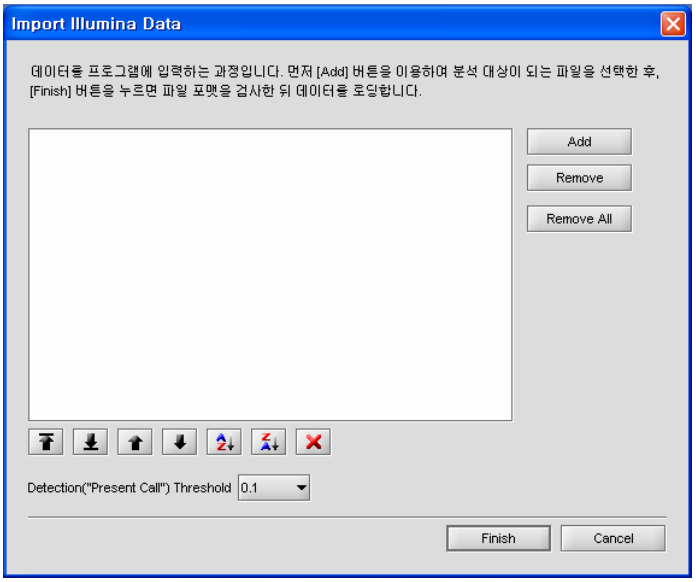
- Others (If no species corresponding)
- All (If species is unknown, browse all species when using annotation function afterwards)

⑥ Click [Next>] button, then the Analysis will be created and the window as below will appear.

⑦ [Click [Add] button to select file to be added, then the file will be added on the list on the left side of the window, and use [Remove] and [Remove All] button to delete the item.

⑧ Detection (“Present Call”) Threshold: This only appears when Illumina chip data is selected and set the range of the Present Call. If you select ‘0.05’, all Probe ID which Detection P-value is under 0.05 will be treated as Present Call, and the rest Probe ID will be processed as Absent Call.

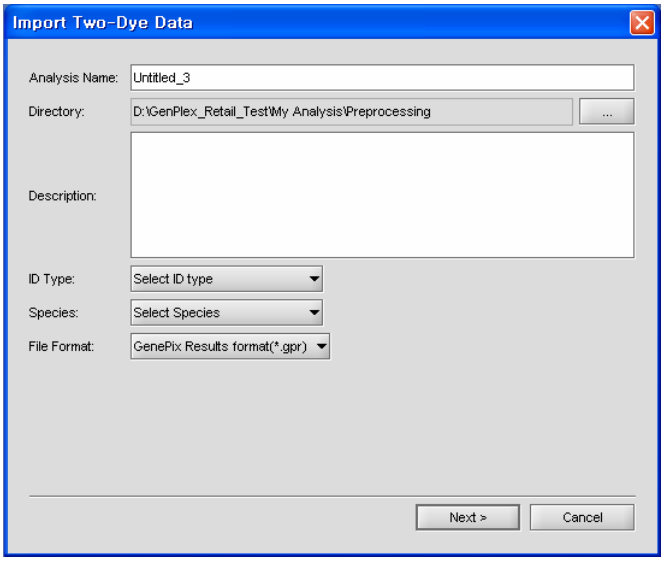
⑨ Click [Finish] button to input the selected data. (go to 2.2 ▷)


Import Illumina Data
데이터를 프로그램에 입력하는 과정입니다. 먼저 [Add] 버튼을 이용하여 분석 대상이 되는 파일을 선택한 후, [Finish] 버튼을 누르면 파일 포맷을 검사한 뒤 데이터를 로딩합니다.
Add
Remove
Remove All
Detection("Present Call") Threshold
0.1
Finish
Cancel

<Figure 2-9> One-Dye Chip Data: Illumina Data Selecting Window

2.1.3. Import Two-dye Chip Data

On the menu bar, select [File] → [Import Two-Dye Chip Data], or click on the third icon , then Analysis Information Input Window appears.



<Figure 2-10> Two-Dye Chip Data: Analysis Data Input Window

① Analysis Name: Input the name of the Analysis File to be created.

② Directory: Click […] button to select location where Analysis file will be created.

③ Description: Input information related on Analysis File (can be skipped).

④ ID Type: Select ID Type of the input data.

- Commercial Product Probe ID
  - Agilent Probe ID(Two-dye)
  - CodeLink Probe ID
  - Illumina Probe ID
  - Operon Probe ID
- Public Database ID
  - IMAGE Clone ID
  - NCBI Clone ID
  - NCBI GenBank Accession
  - NCBI Gene ID (LocusLink)
  - NCBI UniGene ID
- Others (If ID unknown)

⑤ Species: Select type of the data input (Provide different species according to each ID Type)

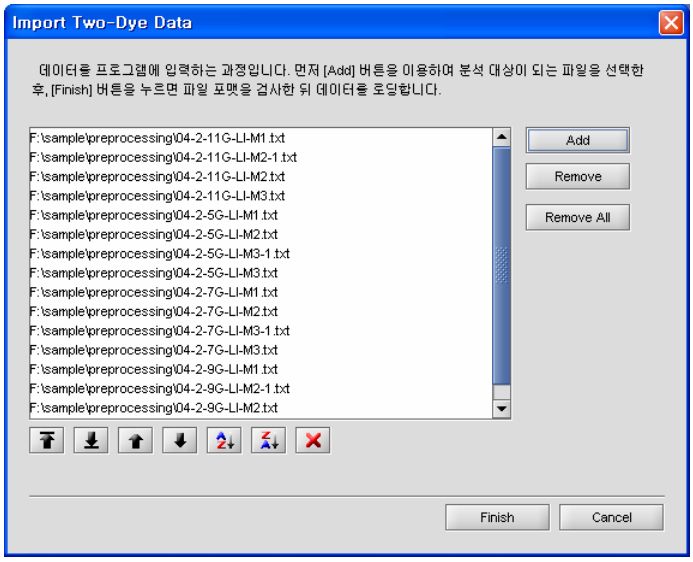
- C.elegans
- Human
- Mouse
- Rat
- Others (If no species corresponding)
- All (If species unknown, browse all kind when using annotation function)

⑥ File Format: Select file format for the input data.

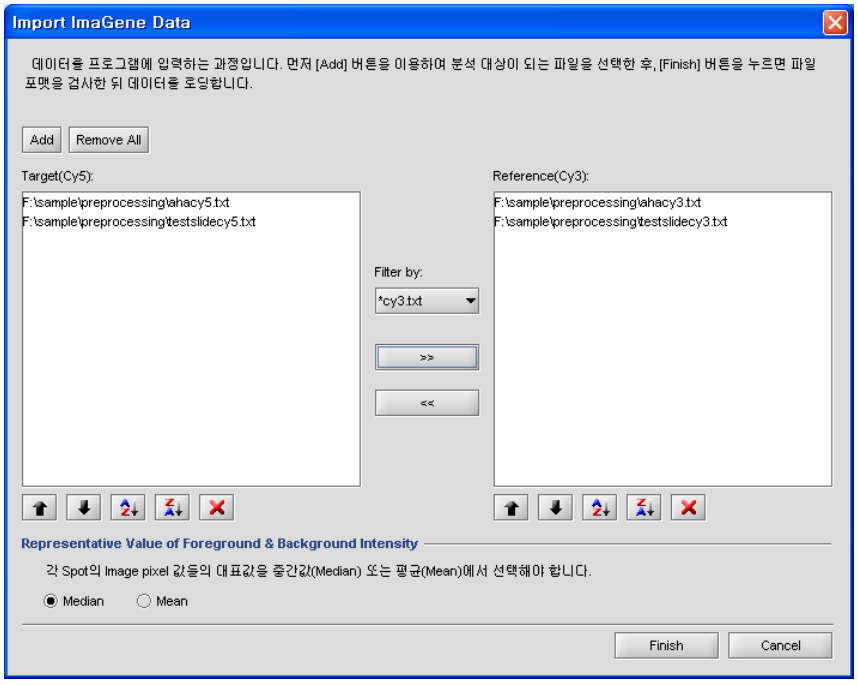
- GenePix Results format(*.gpr)

- ImaGene Data

⑦ Click [Next>] button, then Analysis File will be created and the window appears to select data as seen in the figure below.



<Figure 2-11> Two-Dye Chip Data: Data Selecting Window



<Figure 2-12> Two-Dye Chip Data: Data Selecting Window (ImaGene Data)

⑧ Click [Add] button to select file to input, then the selected file will be added on the left side list. To delete the file, use [Remove] and [Remove All] button. In case of ImaGene Data, designate Cy5 and Cy3 in Pair to input.

⑨ Click [Finish] button to input data selected (go to 2.2 ▷).

- If there is an error in data format.

You can retry after adjusting the error with document editor.

2.1.4. Open Analysis

On the menu bar, select [File] → [Open Analysis], or click on the fourth icon , then you can open saved Analysis File.

2.1.5. Recent Analysis

On the menu bar, select [File] → [Recent Analysis], then you can open the latest analyzed Analysis File. This list can be deleted if you click [Clear History] menu.

2.1.6. Save Analysis

On the menu bar, select [File] → [Save Analysis], or click on the fifth icon , then you can save currently working Analysis File.

2.1.7. Save Analysis As

On the menu bar, select [File] → [Save Analysis As....], or click on the sixth icon , then you can save Analysis file in different name.

2.1.8. Close Analysis

On the menu bar, select [File] → [Close Analysis], then you can close Analysis File.

2.1.9. Analysis Properties

On the menu bar, select [File] → [Analysis Properties], then you can change the attribute of currently working Analysis File.

2.1.10. Configure

On the menu bar, select [File] → [Configure], then you can adjust set screen.

2.1.11. Exit

On the menu bar, select [File] → [Exit], then you can close the program.

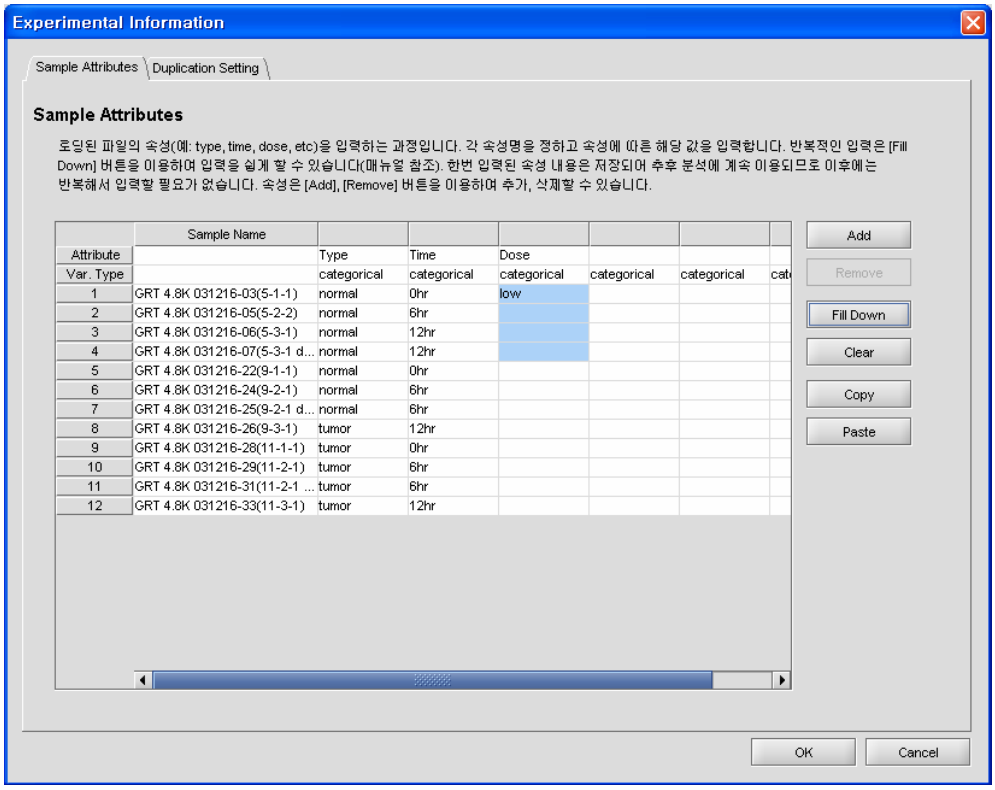
2.2. Preprocessing

2.2.1. Experimental Information

This is the input process of inputting the experiment information of the data. On the menu bar, select [Preprocessing] → [Experimental Information], or click on the seventh icon .

- Sample Attributes

  Input Sample Attributes needed for further analysis. User can classify in different attributes or select data attributes needed only, according to the value that user has input. If no attributes are input, then it may not progress to the next analysis step.


Experimental Information
Sample Attributes
Duplication Setting
Sample Attributes
로딩된 파일의 속성(예: type, time, dose, etc)을 입력하는 과정입니다. 각 속성명을 정하고 속성에 따른 해당 값을 입력합니다. 반복적인 입력은 [Fill Down] 버튼을 이용하여 입력을 쉽게 할 수 있습니다(매뉴얼 참조). 한번 입력된 속성 내용은 저장되어 추후 분석에 계속 이용되므로 이후에는 반복해서 입력할 필요가 없습니다. 속성은 [Add], [Remove] 버튼을 이용하여 추가, 삭제할 수 있습니다.
Add
Remove
Fill Down
Clear
Copy
Paste
OK
Cancel

<Figure 2-13> Sample Attributes Input Window

① Attribute Name: Input Attribute name. Type, Time, Dose are input as basic value. To adjust, use [Add] and [Remove] button to add or delete attributes.

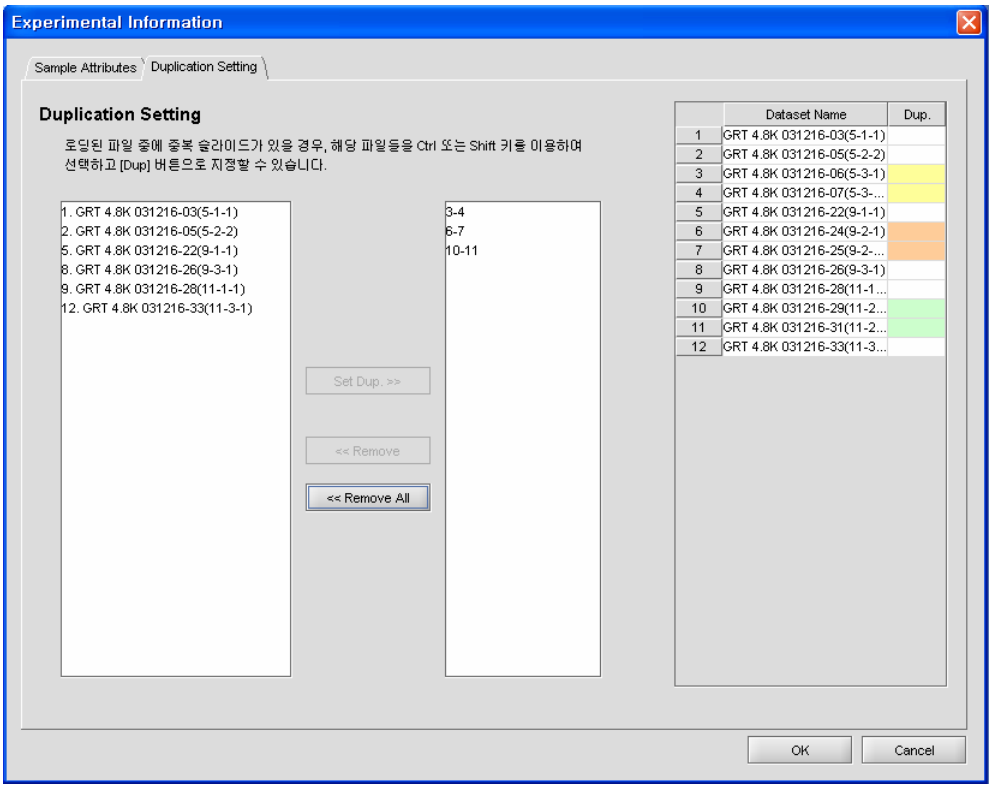
② Var. Type: Select either Categorical or Continuous as attribute type (Currently, supporting only Categorical type).

③ Double click each cell to input attributes, and use [Fill Down], [Copy], [Paste] button for easier input.

④ If there is no duplicate experiment data, click [OK] button to complete. But if duplicate experiment data exists, proceed following Duplication Setting procedure as you can see below.

■ Duplication Setting

Even if the duplicating experiment data exists, set it up. But in case of no duplicate experiment data existing, or in case of Affymetrix GeneChip Data, this procedure does not concern, so this can be skipped.



<Figure 2-14> Duplication Setting Window

① Select duplicate experiment data from the list on the left side of the window (use Ctrl or Shift key) and click [Set Dup.>>] button to duplicate experiment data setting, then it will be added in the list on the right side of the window. Set up data will be shown in the right side figure in same color. To cancel duplicate experiment data, use [Remove] or [Remove All] button.

2.2.2. Filtering Error Spot

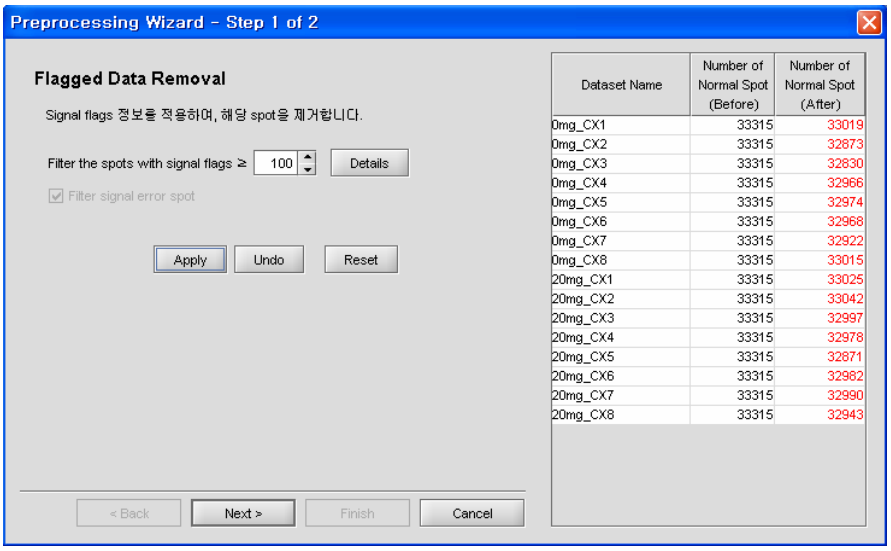
On the menu bar, select [Preprocessing] → [Filtering Error Spot], or just click on the eighth icon , and this corresponds to One-dye Chip Data and Two-dye Chip Data.

2.2.2.1. One-Dye Chip Data

① Flagged Data Removal

This is the procedure to exclude the spot which has bigger value than the user have set up the Flag item value. This method must be applied, otherwise cannot proceed to the following step. Click [Apply] button, then you can see the number of spots before and after applying on

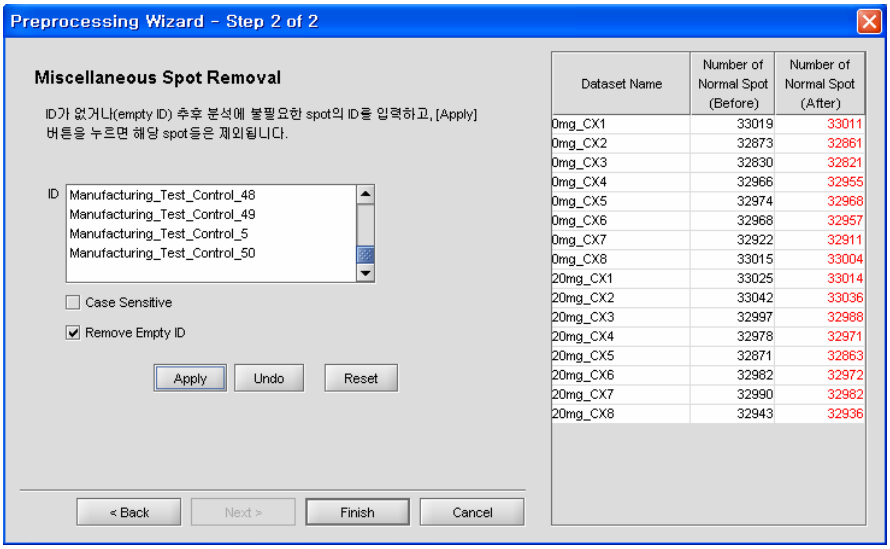
the right side of the table, and click [Next >] button to go to next step.



<Figure 2-15> Flagged Data Removal Set up Window

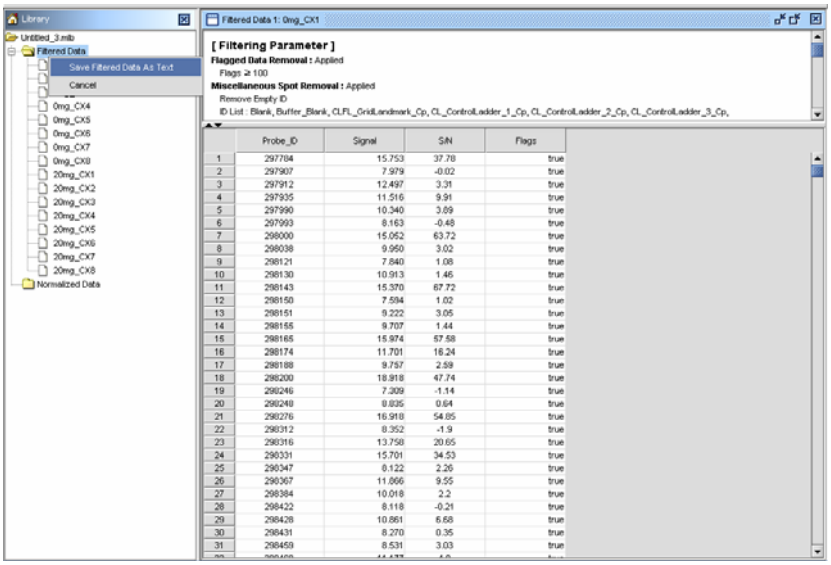
② Miscellaneous Spot Removal

Input unnecessary spot ID list for further analysis, or select option to be deleted, then select [Apply] button, the user can confirm the number of spot before and after applying on the right side list of the window. Click [< Back] button to go to prior step, and click [Finish] and [Cancel] button to complete Filtering Error Spot procedure or delete.



<Figure 2-16> Miscellaneous Spot Removal Set up Window

③ Filtering Result



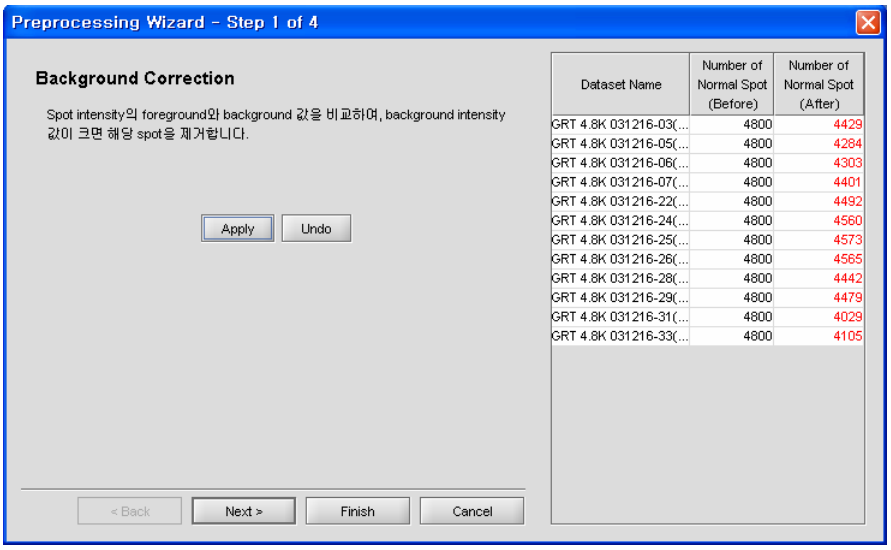
<Figure 2-17> One-Dye Chip Data: Filtering Result Window

- Input data name will be shown on the browse window.
- Double click each data name to confirm each signal spot and related items after applying Filtering.
  - Probe ID: Probe own ID
  - Signal: Value of each probe Signal Intensity
  - S/N: Value of each probe Signal/Noise
  - Flags: Flag data of each probe (True : Valid Spot, False : Filtered Spot)
- Select Before Normalization folder, then right click on the mouse and select [Save Data As Text] menu, then user can save data in text file.

2.2.2.2. Two-Dye Chip Data

① Background Correction

Compare the Background Intensity with Foreground Intensity of the Spot. This is the procedure to exclude the Spot which Background Intensity is higher. This step must be applied, otherwise it cannot proceed to the following step.

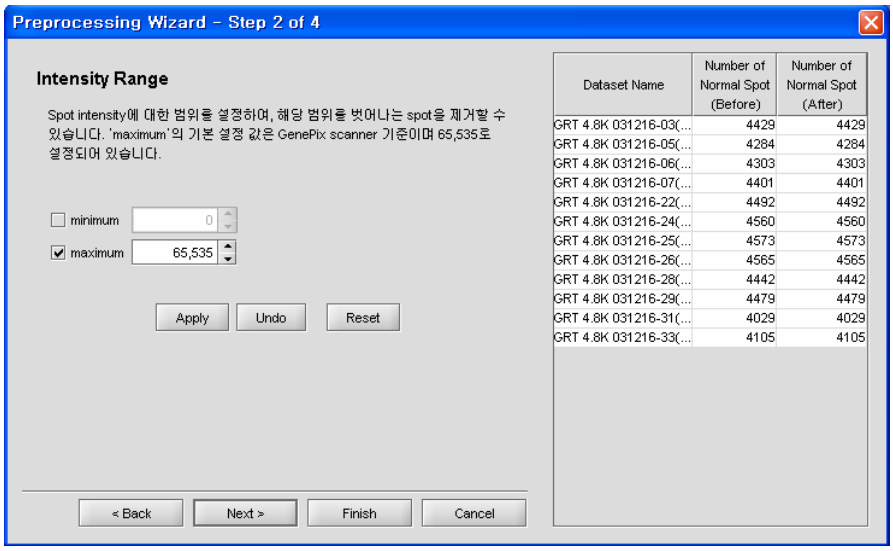

Preprocessing Wizard - Step 1 of 4
Background Correction
Spot intensity의 foreground와 background 값을 비교하여, background intensity 값이 크면 해당 spot을 제거합니다.
Apply
Undo
Dataset Name
Number of Normal Spot (Before)
Number of Normal Spot (After)
GRT 4.8K 031216-03(... 4800 4429
GRT 4.8K 031216-05(... 4800 4284
GRT 4.8K 031216-06(... 4800 4303
GRT 4.8K 031216-07(... 4800 4401
GRT 4.8K 031216-22(... 4800 4492
GRT 4.8K 031216-24(... 4800 4560
GRT 4.8K 031216-25(... 4800 4573
GRT 4.8K 031216-26(... 4800 4565
GRT 4.8K 031216-28(... 4800 4442
GRT 4.8K 031216-29(... 4800 4479
GRT 4.8K 031216-31(... 4800 4029
GRT 4.8K 031216-33(... 4800 4105
< Back
Next >
Finish
Cancel

<Figure 2-18> Background Correction Set up Window

- Click [Apply] button to confirm the number of Spot before and after applying on the right side table of the window.
- Click [Next >]button to proceed to the following step, and use [Finish] and [Cancel] button to complete Filtering Error Spot procedure or delete.

② Intensity Range

Set up the smallest value and the greatest value of Spot intensity, and exclude spot out of this range.


Preprocessing Wizard - Step 2 of 4
Intensity Range
Spot intensity에 대한 범위를 설정하여, 해당 범위를 벗어나는 spot을 제거할 수 있습니다. 'maximum'의 기본 설정 값은 GenePix scanner 기준이며 65,535로 설정되어 있습니다.
minimum
0
maximum
65,535
Apply
Undo
Reset
Dataset Name
Number of Normal Spot (Before)
Number of Normal Spot (After)
GRT 4.8K 031216-03(... 4429 4429
GRT 4.8K 031216-05(... 4284 4284
GRT 4.8K 031216-06(... 4303 4303
GRT 4.8K 031216-07(... 4401 4401
GRT 4.8K 031216-22(... 4492 4492
GRT 4.8K 031216-24(... 4560 4560
GRT 4.8K 031216-25(... 4573 4573
GRT 4.8K 031216-26(... 4565 4565
GRT 4.8K 031216-28(... 4442 4442
GRT 4.8K 031216-29(... 4479 4479
GRT 4.8K 031216-31(... 4029 4029
GRT 4.8K 031216-33(... 4105 4105
< Back
Next >
Finish
Cancel

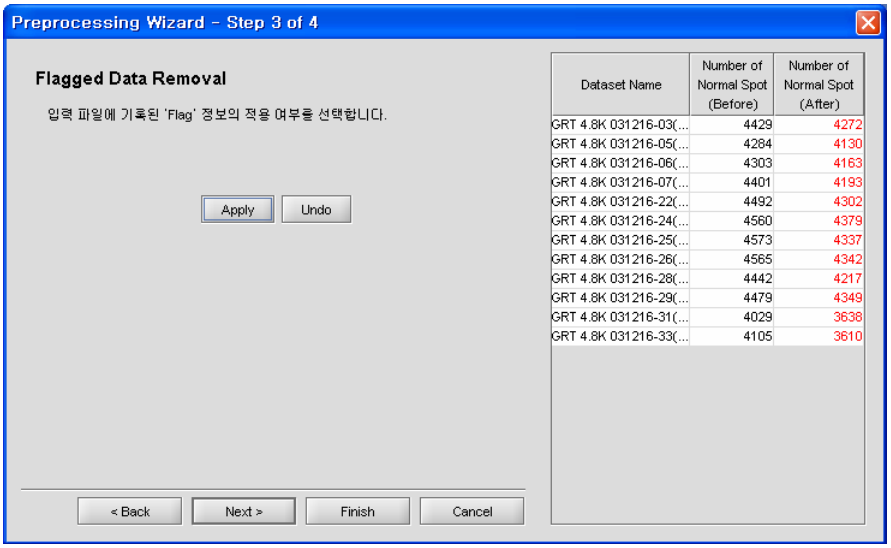
<Figure 2-19> Intensity Range Set up Window

- In Input the smallest value and the greatest value of the Intensity, then click [Apply] button to exclude spot out of this range, and confirm the number of spot before and after applying, on the table right side of the window (Basic Value : Greatest value of GenePix Scanner is 65,535).

- Click [< Back] and [Next >] button to go to previous step or following step, and click [Finish] and [Cancel] button to complete Filtering Error Spot procedure or delete.

③ Flagged Data Removal

Exclude Flagged Spot in the Image Scanner.

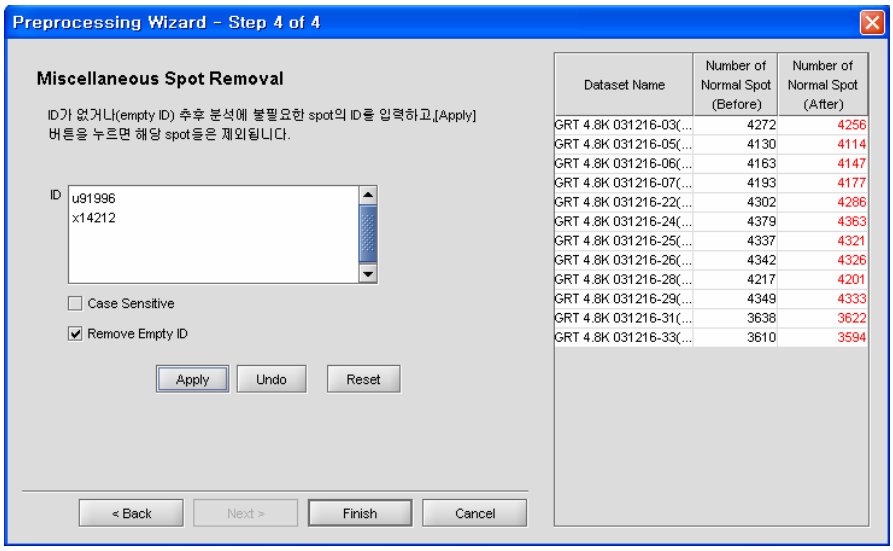


<Figure 2-20> Flagged Data Removal Input Window

- Click [Apply] button to confirm the number of Spot before and after applying on the table right side of the window.
- Click [< Back] and [Next >] button to go to previous step or to the following step, and click [Finish] and [Cancel] button to complete Filtering Error Spot Procedure or delete.

④ Miscellaneous Spot Removal

Possible to exclude unnecessary Spot for further analysis from input data.



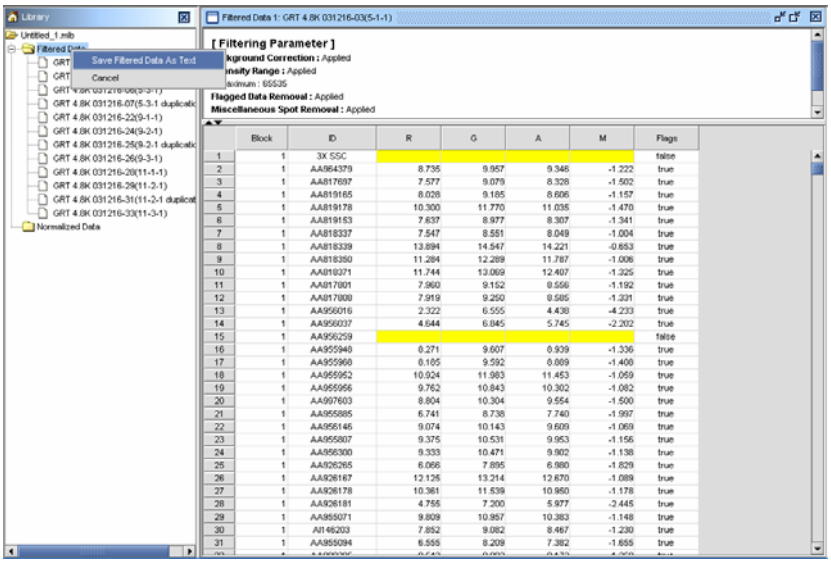
<Figure 2-21> Miscellaneous Spot Removal Input Window

- Input ID list of the unnecessary for further analysis, or select empty ID deleting option,

click [Apply] button to confirm the number of Spot applying before and after on the table right side of the window.

- Click [< Back] button to go to previous step, click [Finish] and [Cancel] button to complete Filtering Error Spot Procedure or delete.

⑤ Filtering Result



<Figure 2-22> Two-Dye Chip Data: Filtering Result Window

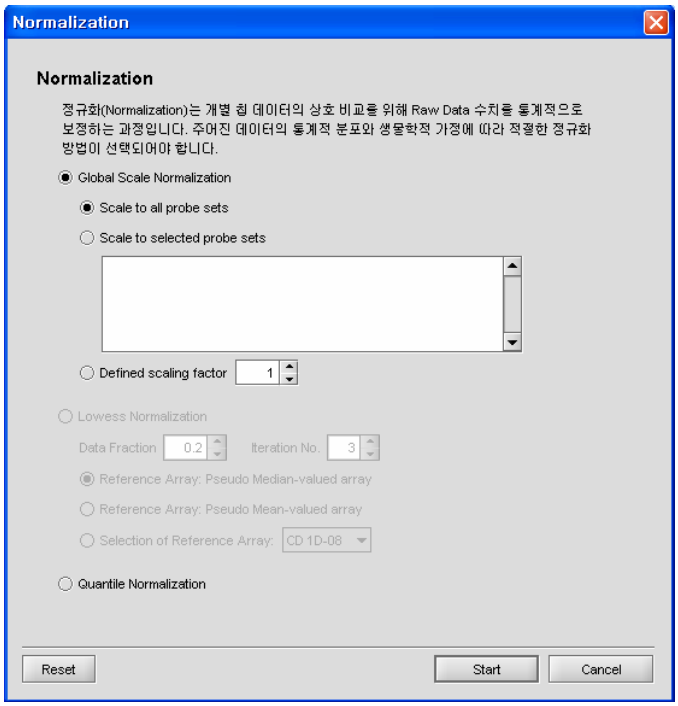
- The name of input data will be shown on the browse window.
- Double click each name of the data to confirm each Intensity spot after applying Filtering and related item.
  - Block: Number of Block of the Spot
  - ID: ID of the Spot
  - R: Intensity Value of Red (treatment) Dye (lo_2 transformed value)
  - G: Intensity Value of Green (control) Dye (log_2 transformed value)
  - A: Average of R and G item
  - M: Ratio of R and G item
  - Flags: Flag Information (true: valid spot, false: filtering spot)
- Select Before Normalization folder, then right click on the mouse and select [Save Data As Text] menu to save data in to text file.

2.2.3. Normalization

On the menu bar, select [Preprocessing] → [Normalization], or click on the ninth icon, then the set up window will appear.

2.2.3.1. Affymetrix Gene Chip Data

Importing Affymetrix Chip Data in [2.1.1], Normalization is proceeded at the same time, so this procedure can be skipped. But it can be used when you change the Normalization method, the new Normalization method will be newly operated overlapping existing Normalization method.



<Figure 2-23> Affymetrix Gene Chip Data: Normalization Setup Window

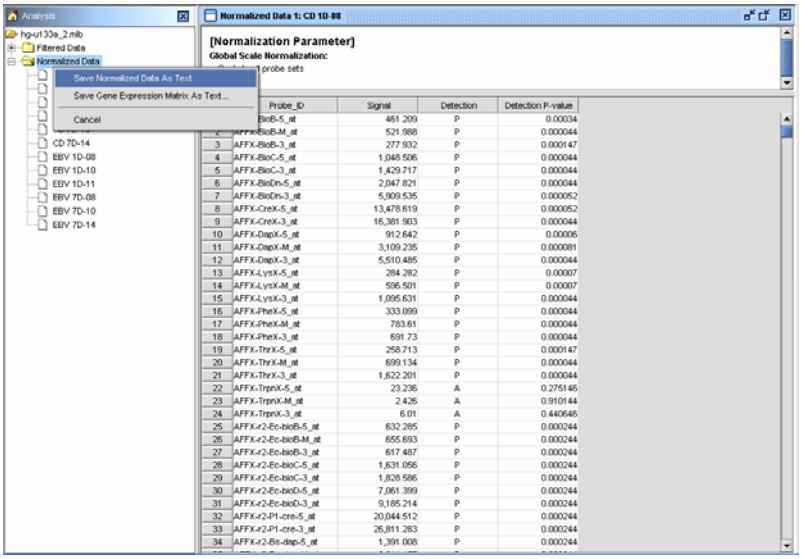
① Global Scale Normalization: This is the method to adjust average signal value of each array according to the option selected as follows [2-2] (It can be selected only when Probe Level Analysis Result is not transformed to $\log_2$ value).
- Scale to all probe sets
- Scale to selected probe sets
- Defined scaling factor

② Lowess Normalization: There is a trend of Lowess Line bending in MA-plot region where intensity range is low or high. Lowess Normalization plays a role to straighten the bended part of the Lowess Line using Local Regression technique [2-3] (This can be selected only when Probe level analysis result is transformed into $\log_2$ value).
- Data Fraction: Possible to set up the data ratio used for the calculation.
- Iteration No: Possible to set up repetition frequency.
- Reference Array: Pseudo Median-valued array
- Reference Array: Pseudo Mean-valued array
- Selection of Reference Array: Possible to set up Reference Array.

③ Quantile Normalization: This is the method to adjust identically all array distributions [2-4].

④ Click [Start] button, then Normalization will be activated.

■ Normalization Result



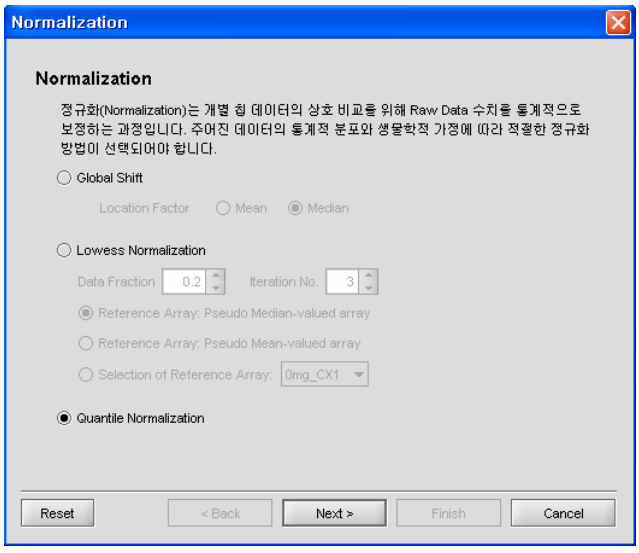
<Figure 2-24> Affymetrix Gene Chip Data: Normalization Result Window

① Normalization Result will be added in the browse window.

② Double click each data name, and after applying Normalization, it is able to confirm Signal of each spot and related items.

③ Right click on the After Normalization folder and click [Save Data As Text] menu to save each data in to text file. Click [Save Gene Expression Matrix As Text…] menu to save data in to GEM format text file.

2.2.3.2. One-Dye Chip Data



<Figure 2-25> One-Dye Chip Data: Normalization Set up Window

① Global Shift

- Mean
- Median

② Lowess Normalization: There is a trend of Lowess Line bending in MA-plot where intensity range is low or high. Lowess Normalization plays a role to straighten the bended part of the Lowess Line using Local Regression technique [2-3].

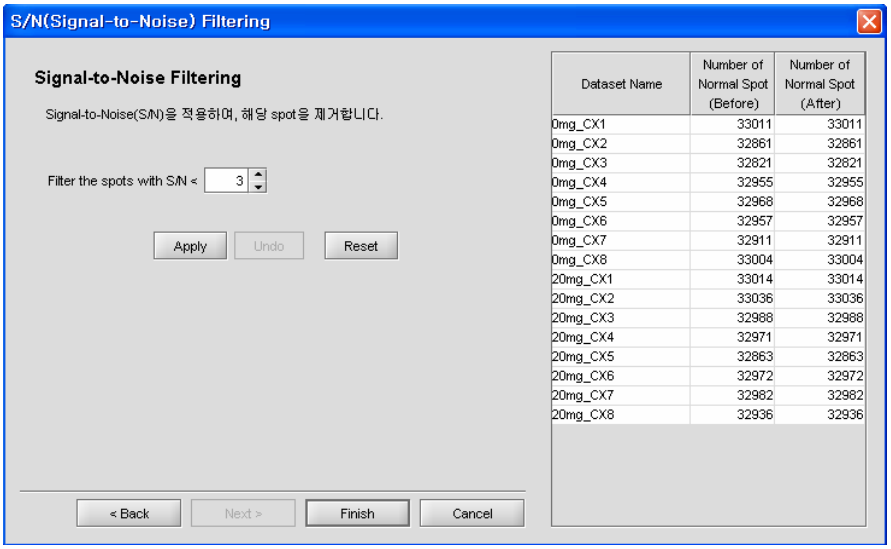
- Data Fraction: Possible to set up the data ratio used for the calculation.
- Iteration No: Possible to set up repetition frequency.
- Reference Array: Pseudo Median-valued array
- Reference Array: Pseudo Mean-valued array
- Selection of Reference Array: Possible to set up Reference Array.

③ Quantile Normalization: This is the method to control array distribution equally [2-4].

④ Click [Next >] button to go to next step.

■ Signal-to-Noise Filtering

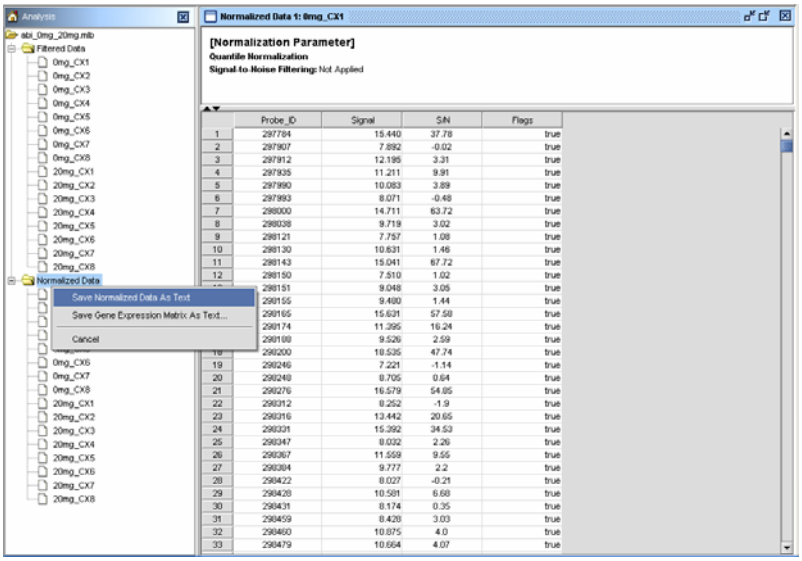
Possible to exclude the spot that has smaller value than the user have set up for the Signal-to-Noise value. This procedure is applied after Normalization is over.



<Figure 2-26> One-Dye Chip Data: Signal-to-Noise Set up Window

① Input the standard value of Signal-to-Noise(S/N) item and click [Apply] button to confirm the number of spot applied before and after, on the table right side of the window. Click [< Back] button to go back to previous step and click [Finish] button to activate Normalization and Filtering.

■ Normalization Result



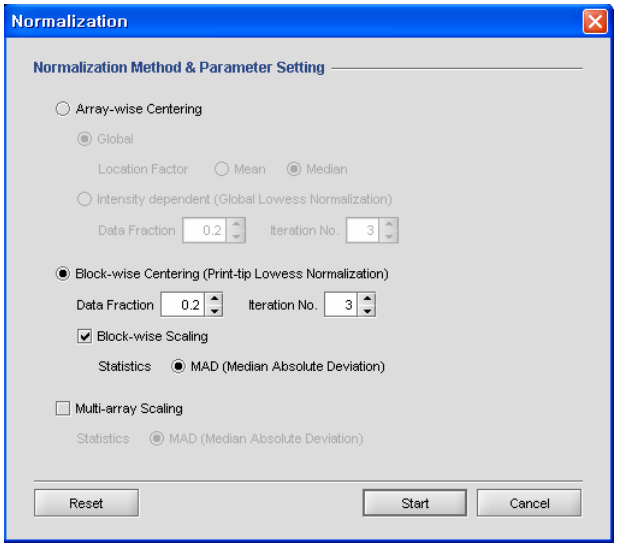
<Figure 2-27> One-Dye Chip Data: Normalization Result Window

① Normalization Result will be added in the browse window.

② Double click each data name, and after applying Normalization, it is possible to confirm Signal of each spot and related items.

③ Right click on the After Normalization folder and click [Save Data As Text] menu to save each data into text file. Click [Save Gene Expression Matrix As Text…] menu to save data into GEM format text file.

2.2.3.3. Two-Dye Chip Data



<Figure 2-28> Two-Dye Chip Data: Normalization Set up Window

① Array-wise Centering : Method to revise classified median value

- Global: Method to correct Mean or Median value.
- Intensity dependent (Global Lowess Normalization): Method to correct using Lowess function [2-3].

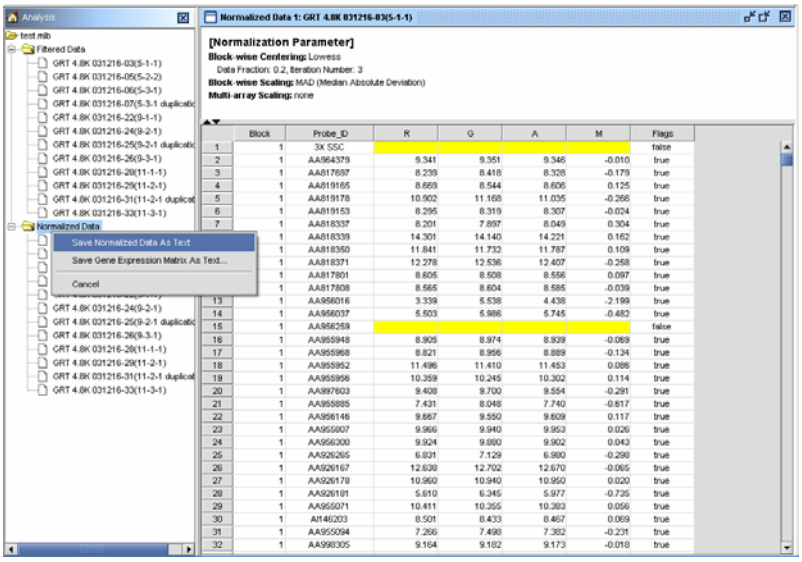
② Block-wise Centering (Print-tip Lowess Normalization): Method to correct using Lowess function to classified block of the slide [2-3].

- Block-wise Scaling: Method to correct with MAD value to classified block scale of the slide.

③ Multi-array Scaling: Method to correct with MAD value the scale of the slide [2-3].

④ Click [Start] button to accomplish Normalization.

■ Normalization Result



<Figure 2-29> Two-Dye Chip Data: Normalization Result Window

① Normalization Result will be added in the browse window.

② Click each data name and after applying Normalization, user can confirm Intensity of each spot and related items.

③ Right click on the After Normalization folder and click [Save Data As Text] menu to save each data into text file. Click [Save Gene Expression Matrix As Text…] menu to save data into GEM format text file.

2.2.4. Set Detection

This concerns only to One-Dye Chip Data, and can set up (change) the Detection Threshold. For the Threshold set up value, double click Nod (Analysis file name) on the top in the browse window,

or can verify selecting [Analysis Properties] on the menu bar.

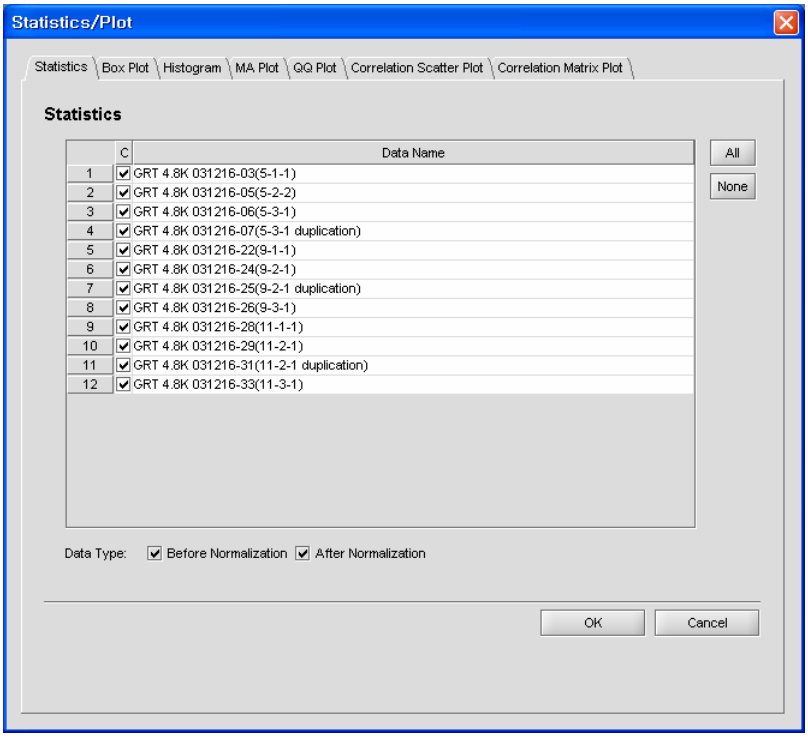
2.2.5. Log Transform

This concerns only to One-Dye Chip Data, and can transform to Log value. But it is not transformable only if it is already transformed, or negative number in Signal.

2.3. Statistics/Plot

Confirm the Statistics of the input data before and after Preprocessing through various Plots. It provides Statistics, Box Plot, Histogram, MA Plot, QQ Plot, Correlation Scatter Plot, Correlation Matrix Plot.

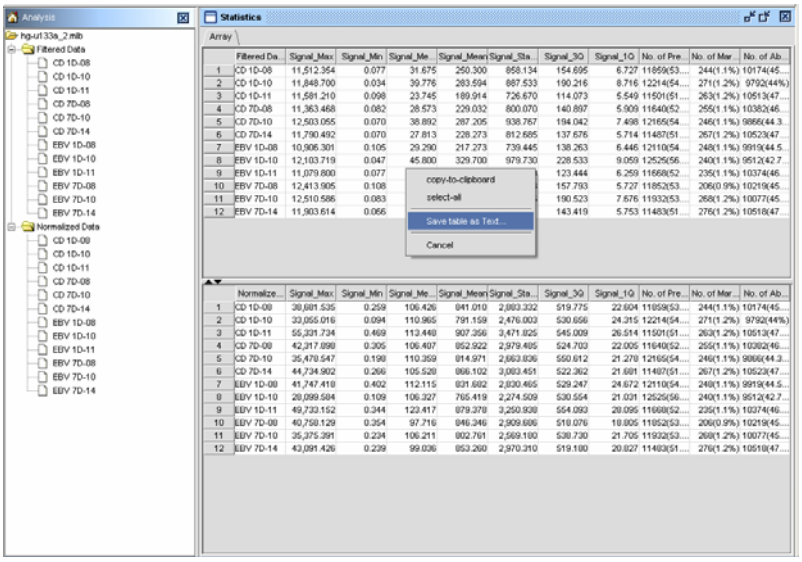
All set up windows are as the figure shown below, first select corresponding tab then select data and data format, click [OK] button to confirm the result.



<Figure 2-30> Statistics/Plot Setup Window

2.3.1. Statistics

It is possible to confirm basic statistic value of input data. On the menu bar, just select [Statistics/Plot] → [Statistics].



<Figure 2-31> Statistics Result Window

① Shows basic statistics of the selected data in table format. Items are as follows.

- Max: Greatest value of the selected data
- Min: Smallest value of the selected data
- Median: Median value of the selected data
- Mean: Average value of the selected data
- Stdev: Standard deviation valued of the selected data
- 3Q: 3rd Quartile of the selected data
- 1Q: 1st Quartile of the selected data

② If a Flag value exists, following item will be shown additionally.

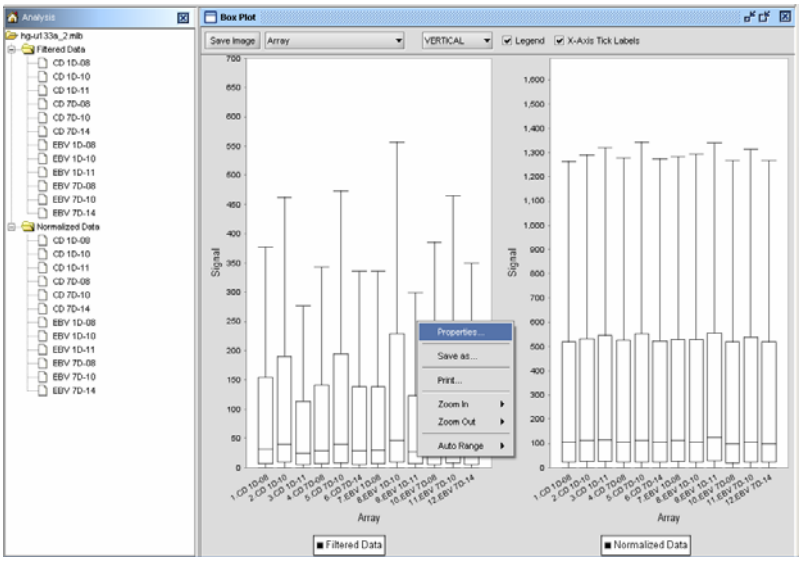
- No. of Flags: Number of false Flag value from the selected data (percentage)

③ In case Affymetrix Gene Chip Data, following items are shown additionally.

- No. of Present Call: Number of P call of each sample (percentage)
- No. of Marginal Call: Number of M call of each sample (percentage)
- No. of Absent Call: Number of A call of each sample (percentage)

2.3.2. Box Plot

On the menu bar, select [Statistics/Plot] → [Box plot], or just click on the tenth icon .

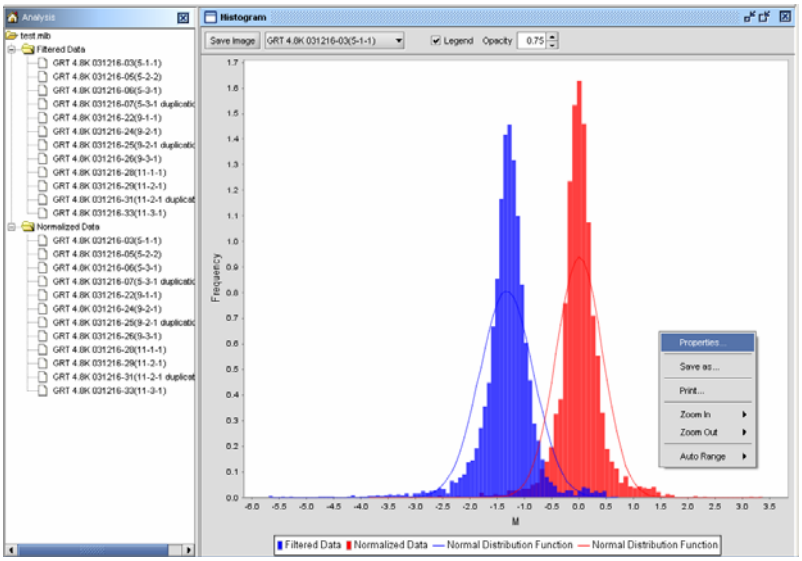


<Figure 2-32> Box Plot Result Window

① Selected data will be shown in a Box Plot, and in case of Two-Dye Chip Data Box Plot of classified Block is supported additionally.

2.3.3. Histogram

On the menu bar, select [Statistics/Plot] → [Histogram], or just click on the eleventh icon .

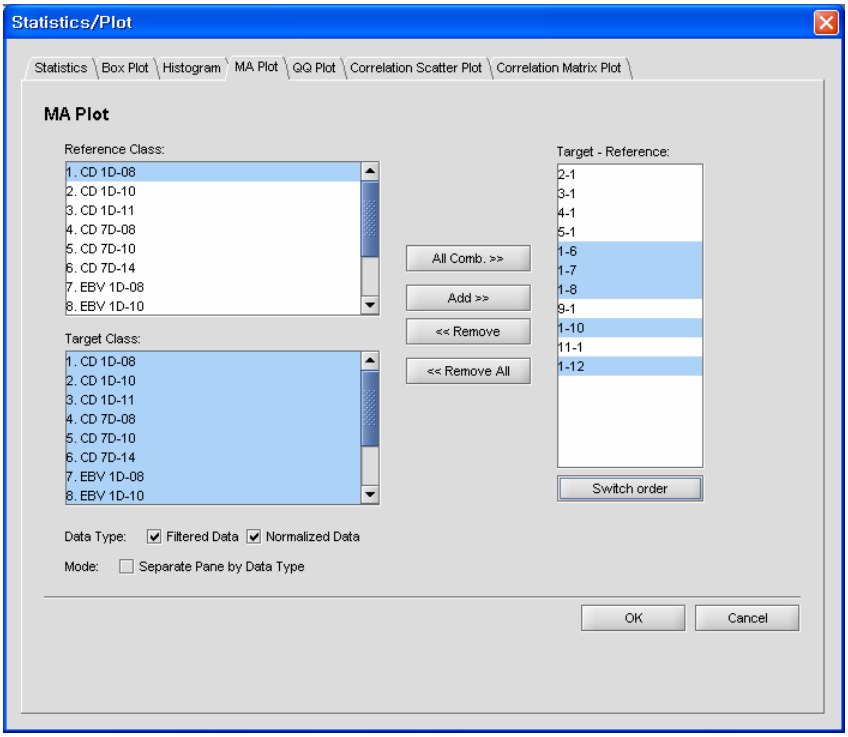


<Figure 2-33> Histogram Result Window

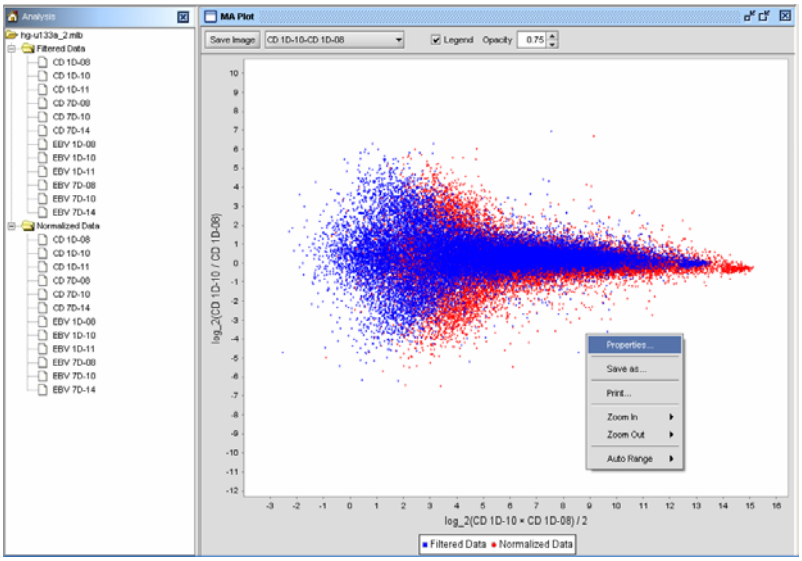
① Can confirm the Histogram of the selected data in classified Array. It is possible to show the Plot, before and after Normalization in a same window or separated window according to set up..

2.3.4. MA Plot

On the menu bar, select [Statistics/Plot] → [MA plot], or just click on the twelfth icon . In case of One-Dye Chip Data, set up window will be shown as below, and designate Reference and Target Class.



<Figure 2-34> One-Dye Chip Data: MA Plot Set up Window

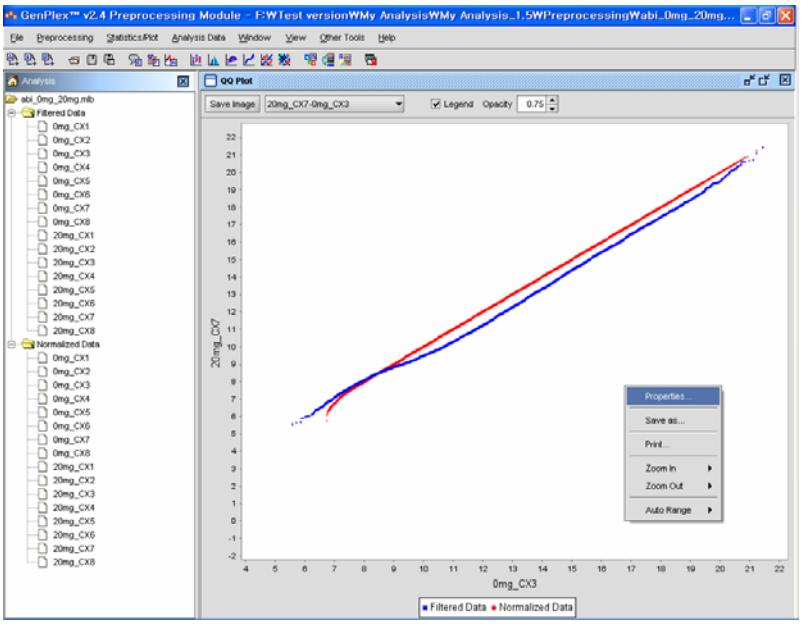


<Figure 2-35> MA plot Result Window

① Can confirm MA Plot in each tab of the selected data, and it is possible to see the Plot before and after Normalization in a same window or separated window according to set up form.

2.3.5. QQ Plot

On the menu bar, select [Statistics/Plot] → [QQ Plot], or just click on the thirteenth icon .

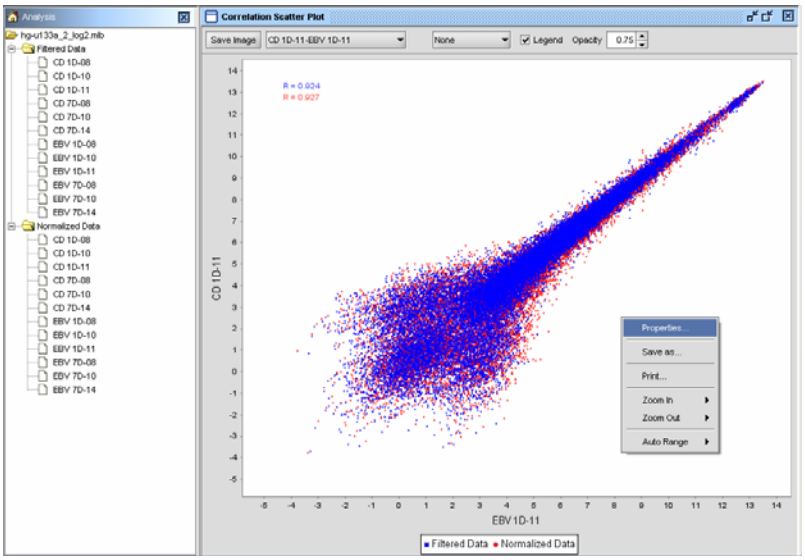


<Figure 2-36> QQ Plot Result Window

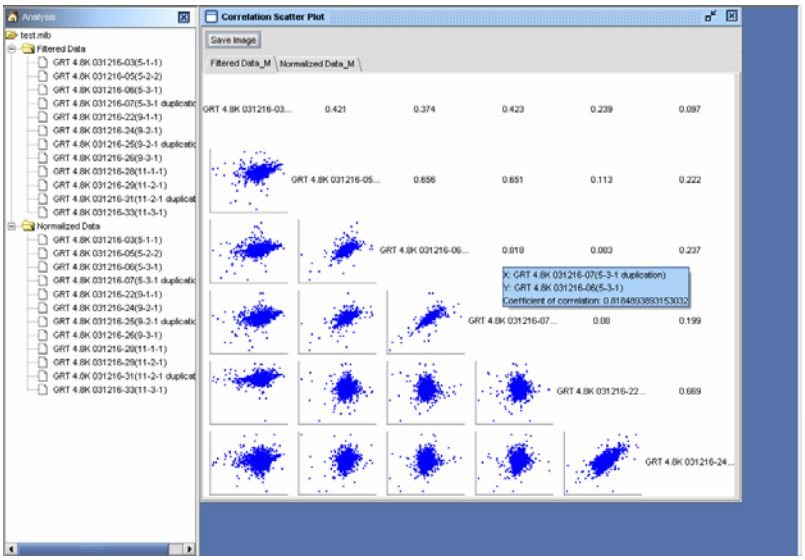
① Can confirm the QQ Plot in each tab of the selected data. It is possible to show the Plot, before and after Normalization in a same window or separated window according to set up.

2.3.6. Correlation Scatter Plot

On the menu bar, select [Statistics/Plot] → [Correlation Scatter Plot], or just click on the fourteenth icon .



<Figure 2-37> Correlation Scatter Plot Result Window 1



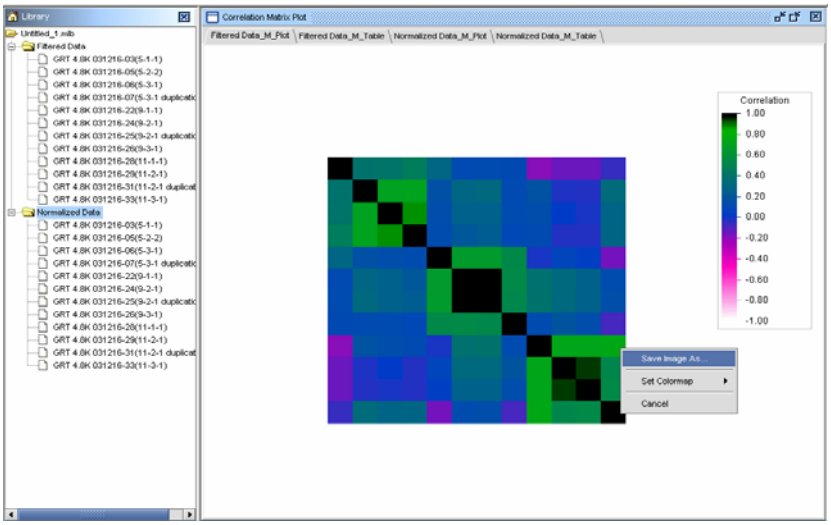
<Figure 2-38> Correlation Scatter Plot Result Window 2

① Can confirm the Correlation Scatter Plot.

② Possible to show before and after Normalization in a same or separated window according to set up, and also possible to show each Plot in one Result Window.

2.3.7. Correlation Matrix Plot

On the menu bar, select [Statistics/Plot] → [Correlation Matrix Plot], or just click on the fifteenth icon .



<Figure 2-39> Correlation Matrix Plot Result Window

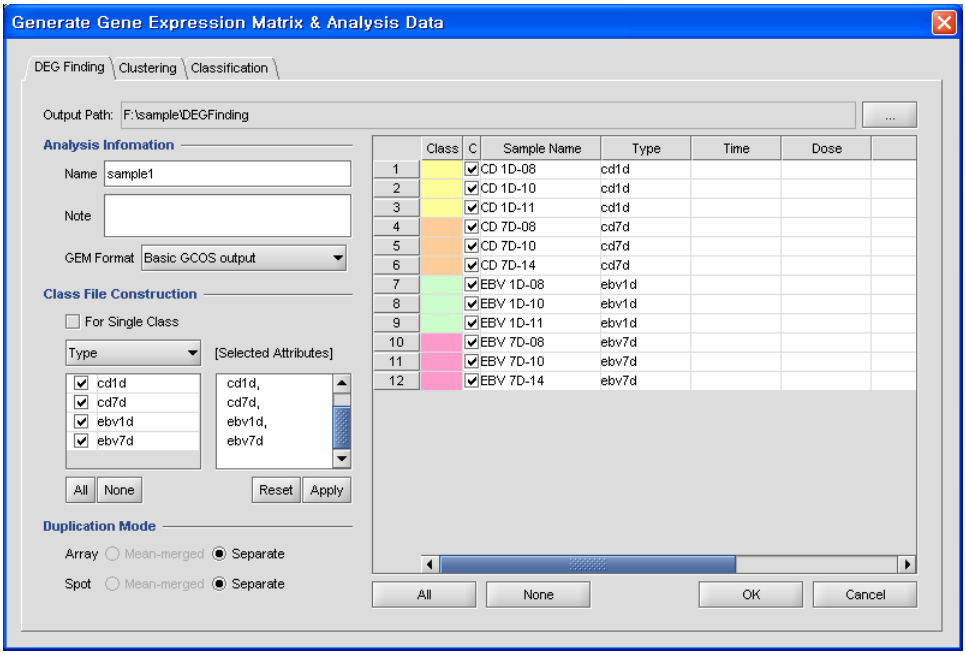
① Can confirm data selected in Correlation Matrix Plot.

2.4. Analysis Data

This is the procedure creating GEM to analyze preprocessed data, and possible to export the created GEM in DEG Finding, Clustering, Classification module.

2.4.1. DEG Finding

This is the DEG Finding module where user can find DEG (Differentially Expressed Gene), can export data. On the menu bar, select [Analysis Data] → [DEG finding], or click on the sixteenth icon to export input data to DEG Finding module.



<Figure 2-40> DEG Finding: GEM Creating Window

① Output Path: Set up pathway for the creating Analysis file.

② Analysis Information

- Name: Input name of the creating Analysis file.
- Note: Input information relating to input data (can skip)
- GEM Format
  - Basic GEM

    This is the basic GEM format, it includes each Array Signal (Intensity).
  - Basic GCOS output

    It corresponds only in case of Affymetrix Gene Chip Data, and GEM is created including Signal, Detection, Detection P-value.
  - GCOS output: Signal+ Detection

    It corresponds only in case of Affymetrix Gene Chip Data, and GEM is created including Signal, Detection

③ Class File Construction

- For Single Class: It corresponds only in case of Two-Dye Data, create selected data in one class.
- Select an attribute: Sample Attribute item set up when inputting data will be shown and will be classified according to the user's choice. Click [Apply] button, then selected data only will be shown on the table right side of the window, and it is shown in different color for each Class which will be very easy to verify.
- Selected Attributes: Show the selected Attribute list.

④ Duplication Mode

Not in case of Affymetrix Gene Chip Data.

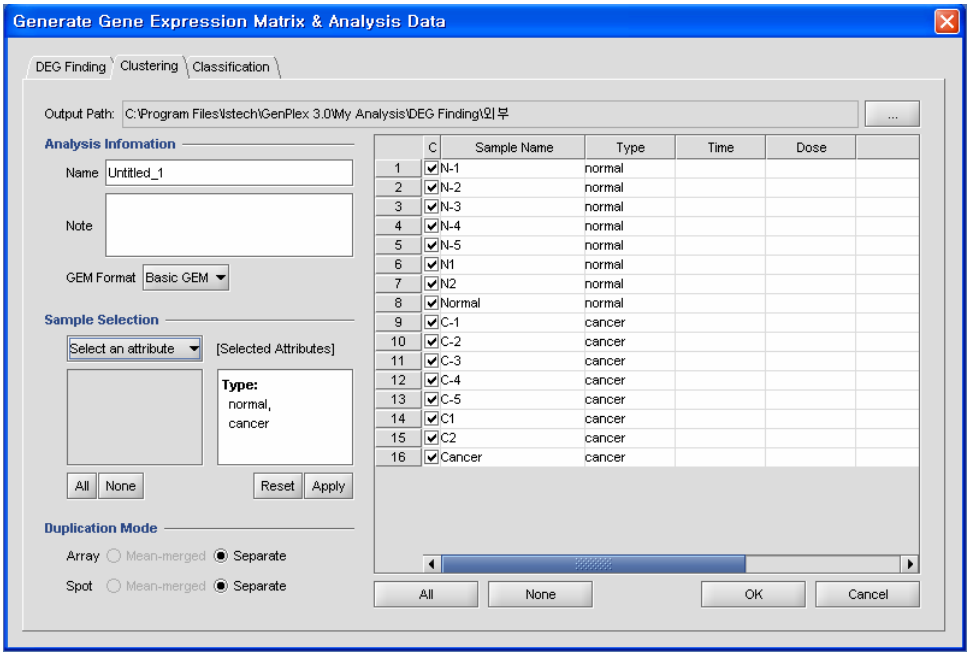
- Array: Select Mean-Merge in case of analyzing average value of the duplication experiment data. If the user does not set up duplication experiment data (Ref.: 2.2.1) Mean-Merge will inactivate.
- Spot: Select Mean-Merge in case of analyzing average value of the same ID Spot within the Array.

⑤ It shows the data only corresponding to the selected Sample Attribute.

⑥ Click [OK] button, the only selected data will be exported and automatically DEG Finding module will activate.

2.4.2. Clustering

Export input data in to Clustering module which is possible for Clustering analysis. On the menu bar, select [Analysis Data] → [Clustering], or click on the seventeenth icon , then it will export the input data to Clustering module.



<Figure 2-41> Clustering: GEM Creating Window

① Output Path: Set up the pathway of Analysis file to be created.

② Analysis Information (Ref.: 2.4.1)

③ Sample Selection

- Select an attribute: Shows Sample Attribute item set up when inputting data, and click [Apply] button, then Sample Attribute selected by the user will only be shown on the table right side of the window.
- Selected Attributes: Shows selected Attribute list.

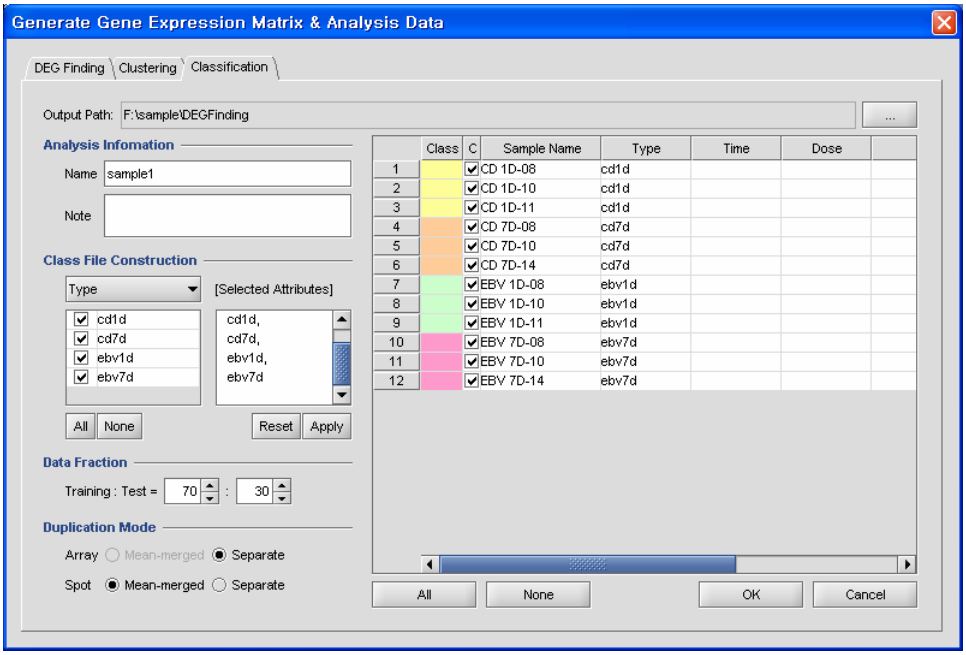
④ Duplication Mode (Ref.: 2.4.1)

⑤ It shows the data, only corresponding to the selected Sample Attribute.

⑥ Click [OK] button to export selected data only and will automatically activate the Clustering module. If Clustering module is already activated, created GEM in working Analysis File will only be added.

2.4.3. Classification

It is possible to Export data to the Classification module used for diagnosis, prognosis and prediction. On the menu bar, select [Analysis Data] → [Classification], or click eighteenth icon to export input data to Clustering module.



<Figure 2-42> Classification: GEM Creating Window

① Output Path: Set up the pathway of the created Analysis file.

② Analysis Information (Ref.: 2.4.1)

③ Class File Construction

- Select an attribute: Sample Attribute item which is set up inputting data is shown and it is classified according to the user's selection. Click [Apply] button, then selected data only

will be shown on the table right side of the window, and it is shown in different colors by Class to easy to distinguish.

- Selected Attributes: Shows selected Attribute list.

④ Data Fraction

The Training Data and the Test Data is composed in random according to the ratio that user has set up based on the number of selected data.

⑤ Duplication Mode (Ref.: 2.4.1)

⑥ It shows the data only corresponding to the selected Sample Attribute.

⑦ Click [OK] button to export selected data and Classification module will automatically activate.

2.5. Reference

[2-1] E. Hubbell, W. Liu, R. Mei (2002) Robust estimators for expression analysis. Bioinformatics, 18(12):1585-1592.

[2-2] Affymetrix (2001) Statistical algorithms reference guide, Technical report, Affymetrix.

[2-3] Y.H. Yang et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30:e15.

[2-4] B.M. Bolstad et al. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2):185-193.

DEG Finding

3. DEG Finding

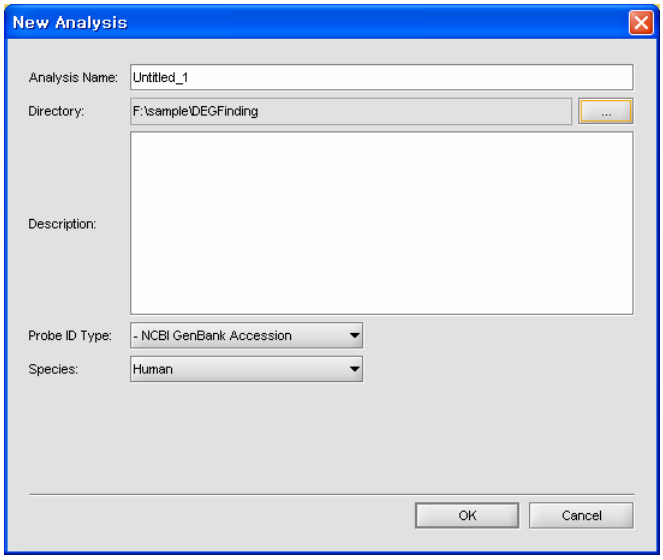
This is the procedure to finding DEG (Differentially Expressed Gene), a procedure to find differentially expressed gene statistically between analysis groups (eg. compare from the reference group to target group).

3.1. File

3.1.1. New Analysis

In Preprocessing module, if the data is exported, the Analysis File will be created automatically, so this procedure does not correspond (go to 3.2 ▷).

On the menu bar, select [File] → [New Analysis], or click on the first icon , then Analysis Creating Window will appear.



<Figure 3-1> Analysis Creating Window

① Analysis Name: Input the name of the Analysis File to be created.

② Directory: Click […] button to select where Analysis File will be created.

③ Description: Input additional information related to the Analysis File (it can be skipped)

④ Probe ID Type: Select ID Type to be input.

- Commercial Product Probe ID
  - Affymetrix GeneChip Probe ID
  - Agilent Probe ID(One Dye)
  - Agilent Probe ID(Two Dye)

- Applied Biosystems 1700 Probe ID
- CodeLink Probe ID
- Illumina Probe ID
- Operon Probe ID

● Public DataBase ID
- IMAGE Clone ID
- NCBI Clone ID
- NCBI GenBank Accession
- NCBI GeneID (LocusLink)
- NCBI UniGene ID

● Others (If ID is unknown)

⑤ Species: Select the species of the data to be input.

⑥ Click [OK] button to create Analysis, then the Data Selecting Window will appear (go to 3.1.7 ▷).

3.1.2. Open Analysis

On the menu bar, select [File] → [Open Analysis], or click on the second icon to open saved Analysis.

3.1.3. Recent Analysis

On the menu bar, select [File] → [Recent Analysis] to open updated analyzed Analysis. This list can be deleted using [Clear History] menu.

3.1.4. Save Analysis

On the menu bar, select [File] → [Save Analysis], or click on the third icon to save Analysis under operation.

3.1.5. Save Analysis As…

On the menu bar, select [File] → [Save Analysis As...], or click on the fourth icon to save Analysis under operation in different name.

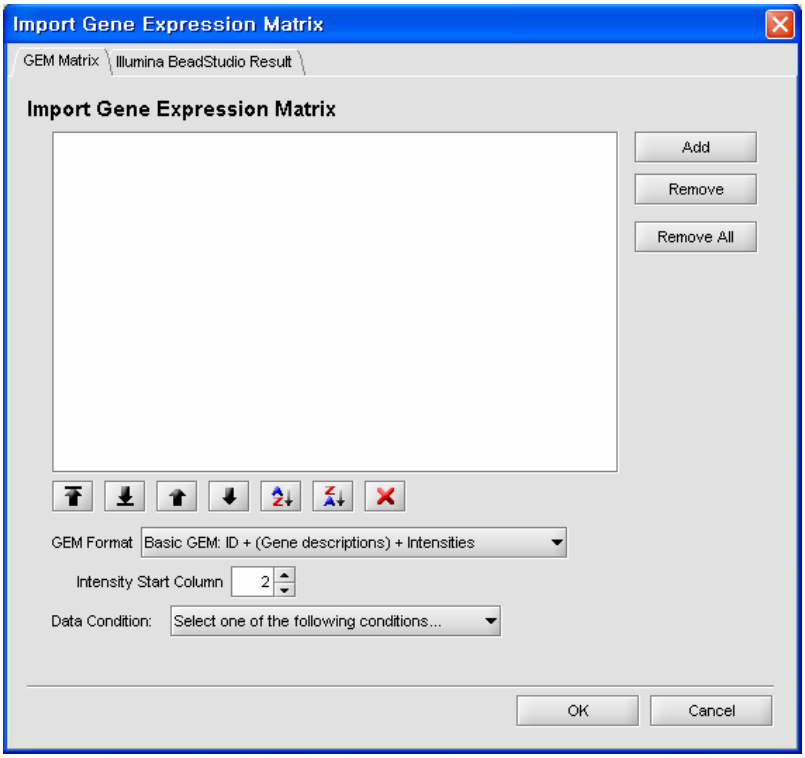
3.1.6. Close Analysis

On the menu bar, select [File] → [Close Analysis] to close Analysis under operation.

3.1.7. Import Data

On the menu bar, select [File] → [Import Data], or click on the fifth icon , then Data Selecting Window will appear.

■ GEM Matrix



<Figure 3-2> Data Selecting Window

① Click [Add] button to select the file to be input, then it will be added in the list on left side of the window, and use [Remove] and [Remove All] button to delete item.

② GEM Format

- Basic GEM: ID+ (Gene description)+ Intensities
  - Intensity Start column: In case of Basic GEM, user can set up the position where the Intensity Column will be started. You can use the Description information in case Description Column exists between ID Column and Intensity.

- Basic GCOS output: Signal+ Detection+ Detection p-value
- Basic GCOS output: Signal+ Detection

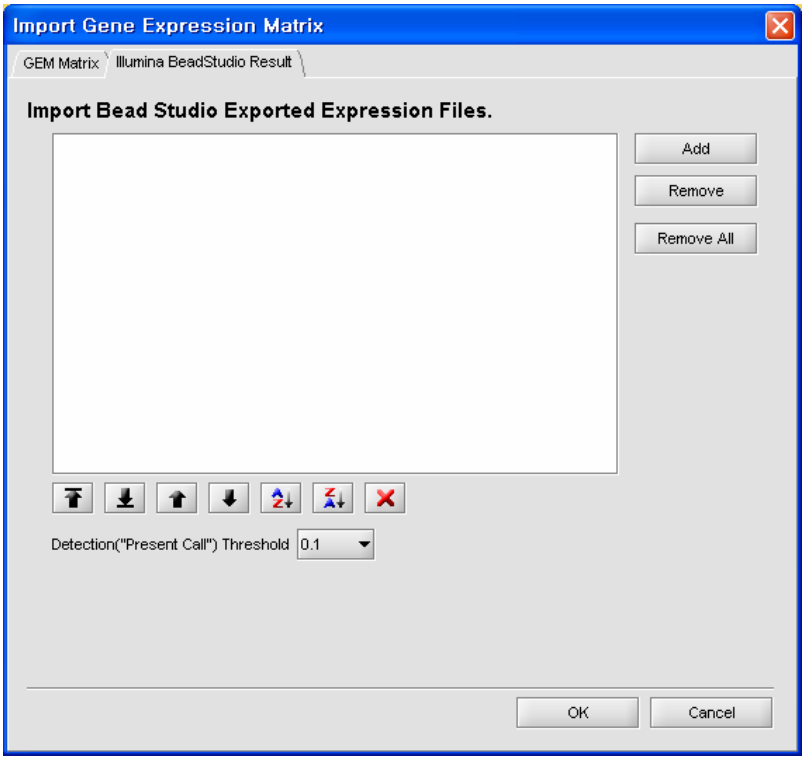
③ Data condition: Select the Log transformed data which will be input.

- The data was not log-transfomed
- The data was log-transformed with base 2
- The data was log-transformed with base 10
- The data was log-transformed with base e

④ Click [OK] button to input data.

■ Illumina BeadStudio Result

It is able to use Illumina file from BeadStudio to input into DEG Finding module, and create Detection column setting up the Threshold value of Detection P-value within the file.



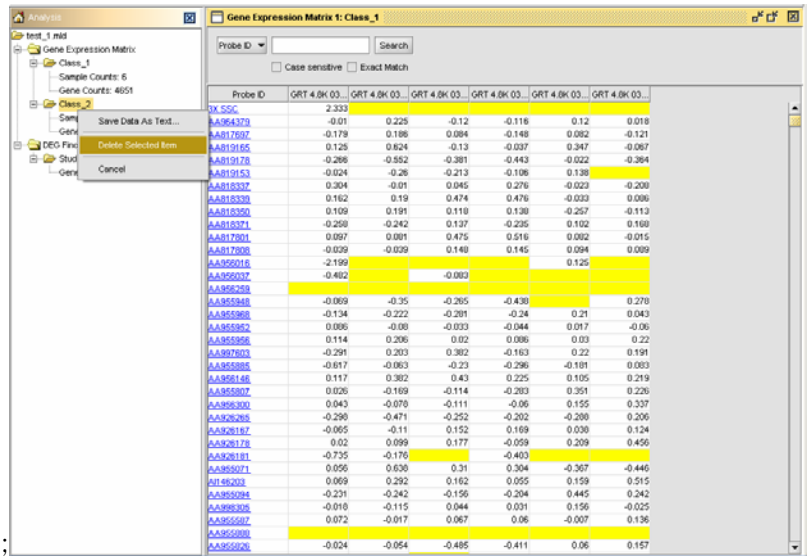
<Figure 3-3> Illumina Data Selecting Window

① Click [Add] button to select file to be input, then the list will be added in the list on the left side of the window. Use [Remove] and [Remove All] button to delete item.

② Detection("Present Call") Threshold

- There is no Detection column as “Signal+ Detection P-value” as file from BeadStudio. Create Detection column setting up the Threshold value of the Detection P-value, then it is possible to delete with Detection Call from the DEG Filtering method when selecting DEG.

■ Data Input Result



<Figure 3-4> Input Data Verifying Window

- To confirm input data, double click on the name of the data in the browse window. Missing Value will be marked in Yellow.

3.1.8. Analysis Properties

On the menu bar, select [File] → [Analysis Properties] to adjust Analysis information under operation.

3.1.9. Exit

On the menu bar, select [File] → [Exit] to close the program.

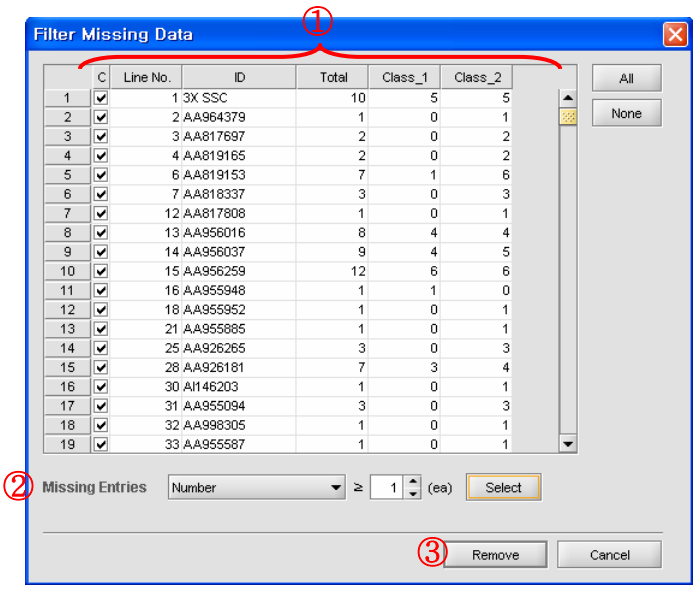
3.2. Preprocessing

3.2.1. Check & Match Data

On the menu bar, select [Preprocessing] → [Check & Match Data] to confirm whether the number of gene of inputting data matches with the ID. If it does not match, Analysis cannot be processed.

3.2.2. Filter Missing Data

On the menu bar, select [Preprocessing] → [Filter Missing Data], then the window as you can see below appears, and it is possible to delete Missing Entry from the input data.



<Figure 3-5> Missing Data Delete Window

① Shows genes' information with Missing value in table format.

- Line No: Order of genes from the input data
- ID: ID of each gene
- Total: Total of missing value of each Class
- Name of each Class: Number of missing value of each Class

② Missing Entries: Select the standard and click [Select] button to properly select genes to be deleted.

- Number: Select genes to be deleted based on the Total Number (column) of missing value.
- Total Percentage: Select genes to be deleted based on the ratio of Total

Number of Samples and Total Number (column) of missing value.

- Class-specific Percentage: Select genes to be deleted based on the ratio of the missing value in each class.

③ Click [Remove] button to delete Missing Entry.

3.2.3. Impute Data

On the menu bar, select [Preprocessing] → [Impute Data], then it is able to complete Missing Value according to the regular rule.

3.2.4. Log Transform

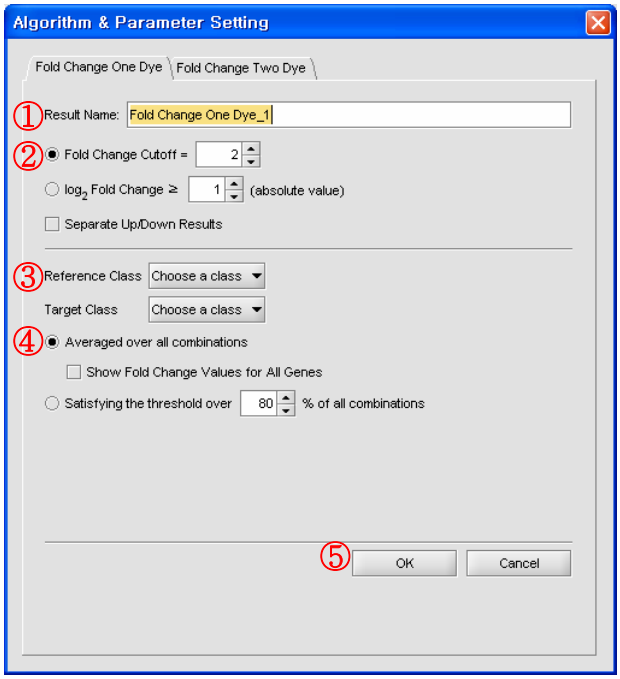
On the menu bar, select [Preprocessing] → [Log Transform] to transform input data into Log value.

3.3. DEG Finding

3.3.1. Fold Change

3.3.1.1. Fold Change One Dye

On the menu bar, select [DEG Finding] → [Fold Change] → [Fold Change One Dye], then set up window appears.


Algorithm & Parameter Setting
Fold Change One Dye
Fold Change Two Dye
① Result Name: Fold Change One Dye_1
② Fold Change Cutoff = 2
log2 Fold Change ≥ 1 (absolute value)
Separate Up/Down Results
③ Reference Class Choose a class
Target Class Choose a class
④ Averaged over all combinations
Show Fold Change Values for All Genes
Satisfying the threshold over 80 % of all combinations
⑤ OK
Cancel

<Figure 3-6> Fold Change One Dye Set Up Window

① Result Name: Input the name of the result to be created.

② Select the cut off to be applied.

- Fold Change Cutoff: Value of Fold Change which is not transformed to Log value (Basic value : 2)
- log_2 Fold Change: Value of Fold Change which is transformed to Log_2 value (Basic value : 1)
- Separate Up/Down Results: Each up-regulation gene and down-regulation gene list will be created based on the Fold Change result.

③ Select reference class and target class of the applying data.

④ Select the method to be applied.

- Average over all combinations: Genes with more than average cut off of all Cold Change combination of reference class and target class will be selected.
  - Show Fold Change Values for All Genes: Shows the Fold Change Value of

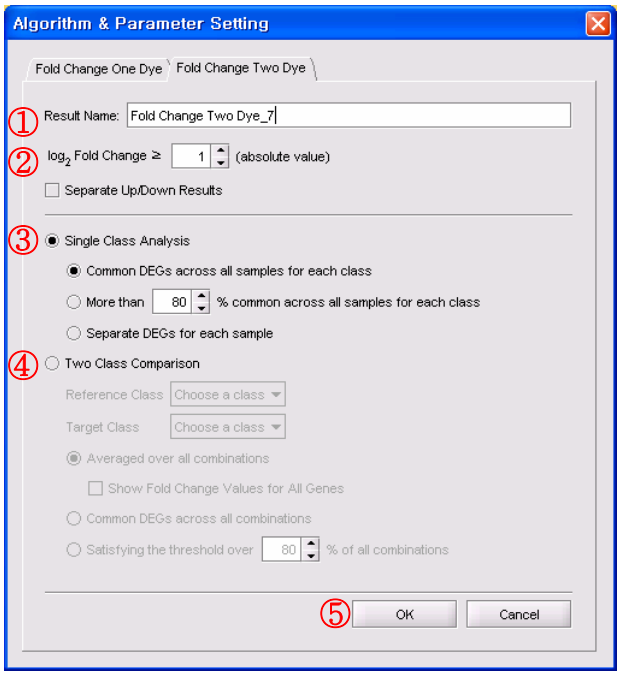
all genes. If [Separate Up/Down Results] option is selected as seen above, this function will not activate.

- Satisfying the threshold over □% of all combinations: Genes with more than ratio that user have set up from all Fold Change Combination.

⑤ Click [OK] button to see the result window.

3.3.1.2. Fold change Two Dye

On the menu bar, select [DEG finding] → [Fold change] ->[Fold Change Two Dye], then set up window appears.


Algorithm & Parameter Setting
Fold Change One Dye
Fold Change Two Dye
① Result Name: Fold Change Two Dye_7
② log2 Fold Change ≥ 1 (absolute value)
Separate Up/Down Results
③ Single Class Analysis
Common DEGs across all samples for each class
More than 80 % common across all samples for each class
Separate DEGs for each sample
④ Two Class Comparison
Reference Class Choose a class
Target Class Choose a class
Averaged over all combinations
Show Fold Change Values for All Genes
Common DEGs across all combinations
Satisfying the threshold over 80 % of all combinations
⑤ OK
Cancel

<Figure 3-7> Fold Change Two Dye Set Up Window

① Result Name: Input the name of the Result which will be created.

② Select the cut off to be applied.

- log_2 Fold Change: Value of the Fold Change which is Log_2 transformed (Basic value : 1)
- Separate Up/Down Results: Each up-regulation genes and down-regulation genes' list will be created resulting from Fold Change.

③ Single Class Analysis: It is applied when significant genes are selected from each single class.

- Common DEGs across all samples: The genes' intensity that is above cut off of all samples from each class, is selected.

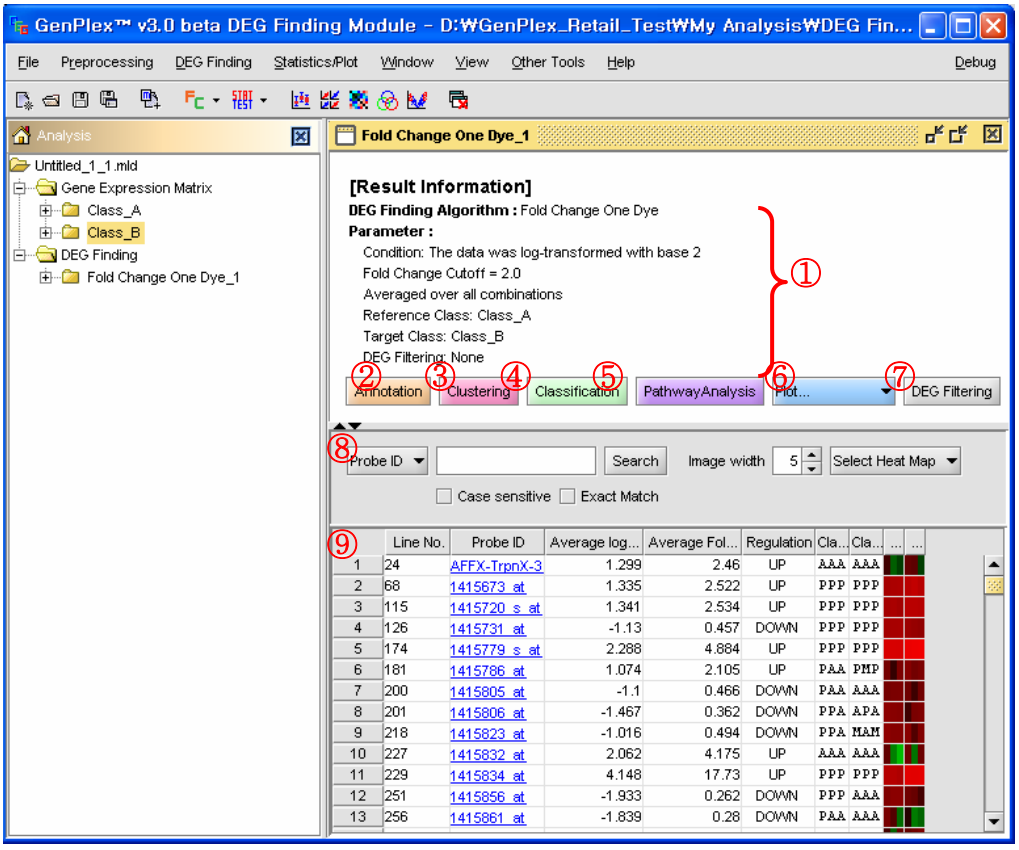
- More than □% common across all samples: The genes' ratio that user have set up from all sample is over the cut off, is selected.
- Separate DEGs each sample: The genes over each sample's cut off are selected. If this option is selected [Separate Up/Down Results] function as seen above does not activate.

④ Two Class Comparison: It is applied when comparing Two Classes. First select reference class and target class of the data and then select the method to be applied (Ref.: 3.3.1.1.)

- Common DEGs across all combinations: The genes of all Fold Change combinations over cut off are selected.

⑤ Click [OK] button, then the result window will appear.

3.3.1.3. Fold Change

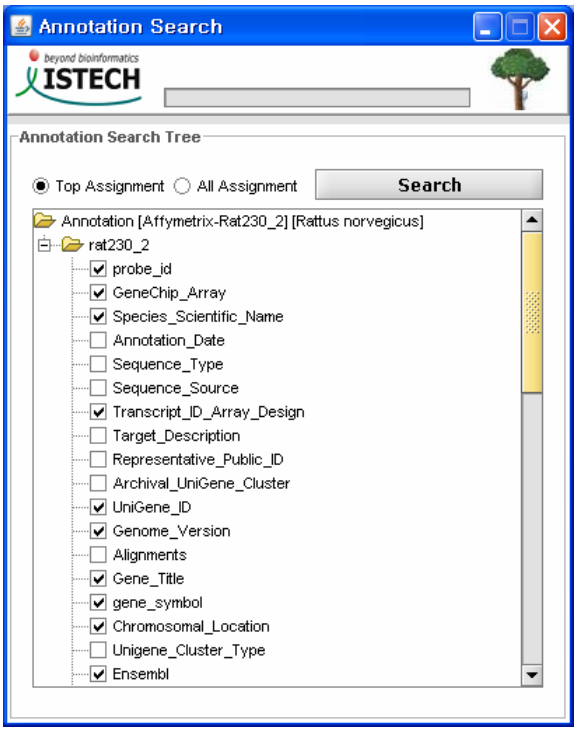


<Figure 3-8> Fold Change Result Window

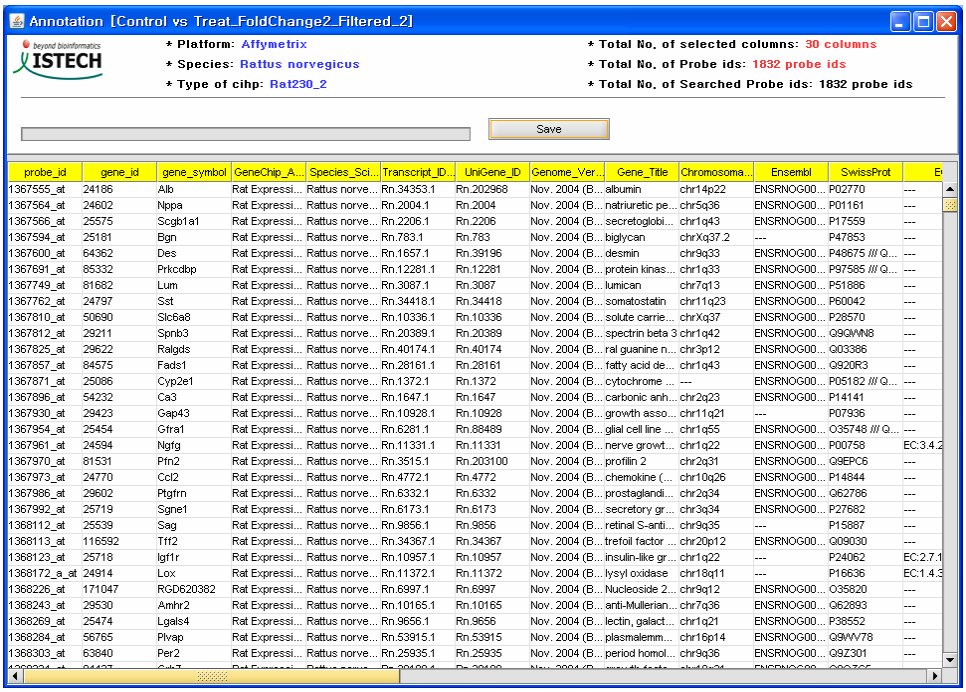
① Shows the DEG Finding Algorithm and Parameter set up which user have selected.

② Annotation

- Annotation: It shows all Annotation Information in one table.
- Annotation Search Tree
  - Top Assignment: Search top assigned information setting up as default
  - All Assignment: Search all assigned information
- Can identify the information on Chip Type and Species as seen on the search window.
- Click Search to search Annotation.
- In case of Commercial Platform, it includes all Annotation information provided from each Platform, but in case of KEGG and Uniprot information is added.



<Figure 3-9> Annotation Search Window


Annotation [Control vs Treat_FoldChange2_Filtered_2]
* Platform: Affymetrix
* Species: Rattus norvegicus
* Type of chip: Rat230_2
* Total No. of selected columns: 30 columns
* Total No. of Probe ids: 1832 probe ids
* Total No. of Searched Probe ids: 1832 probe ids
Save
probe_id gene_id gene_symbol GeneChip_A... Species_Sci... Transcript_ID... UniGene_ID Genome_Ver... Gene_Title Chromosoma... Ensembl SwissProt
1367555_at 24186 Alb Rat Expressi... Rattus norve... Rn.34353.1 Rn.202968 Nov. 2004 (B... albumin chr14p22 ENSRNOG00... P02770
1367564_at 24602 Nppa Rat Expressi... Rattus norve... Rn.2004.1 Rn.2004 Nov. 2004 (B... natriuretic pe... chr5q36 ENSRNOG00... P01161
1367566_at 25575 Scgb1a1 Rat Expressi... Rattus norve... Rn.2206.1 Rn.2206 Nov. 2004 (B... secretoglobi... chr1q43 ENSRNOG00... P17559
1367594_at 25181 Bgn Rat Expressi... Rattus norve... Rn.783.1 Rn.783 Nov. 2004 (B... biglycan chrXq37.2 --- P47853
1367600_at 64362 Des Rat Expressi... Rattus norve... Rn.1657.1 Rn.39196 Nov. 2004 (B... desmin chr9q33 ENSRNOG00... P48675 /// Q...
1367691_at 85332 Prkcdbp Rat Expressi... Rattus norve... Rn.12281.1 Rn.12281 Nov. 2004 (B... protein kinas... chr1q33 ENSRNOG00... P97585 /// Q...
1367749_at 81682 Lum Rat Expressi... Rattus norve... Rn.3087.1 Rn.3087 Nov. 2004 (B... lumican chr7q13 ENSRNOG00... P51886
1367762_at 24797 Sst Rat Expressi... Rattus norve... Rn.34418.1 Rn.34418 Nov. 2004 (B... somatostatin chr11q23 ENSRNOG00... P60042
1367810_at 50690 Slc6a8 Rat Expressi... Rattus norve... Rn.10336.1 Rn.10336 Nov. 2004 (B... solute carrie... chrXq37 ENSRNOG00... P28570
1367812_at 29211 Spnb3 Rat Expressi... Rattus norve... Rn.20389.1 Rn.20389 Nov. 2004 (B... spectrin beta 3 chr1q42 ENSRNOG00... Q9QWN8
1367825_at 29622 Ralgds Rat Expressi... Rattus norve... Rn.40174.1 Rn.40174 Nov. 2004 (B... ral guanine n... chr3p12 ENSRNOG00... Q03386
1367857_at 84575 Fads1 Rat Expressi... Rattus norve... Rn.28161.1 Rn.28161 Nov. 2004 (B... fatty acid de... chr1q43 ENSRNOG00... Q920R3
1367871_at 25086 Cyp2e1 Rat Expressi... Rattus norve... Rn.1372.1 Rn.1372 Nov. 2004 (B... cytochrome ... --- ENSRNOG00... P05182 /// Q...
1367896_at 54232 Ca3 Rat Expressi... Rattus norve... Rn.1647.1 Rn.1647 Nov. 2004 (B... carbonic anh... chr2q23 ENSRNOG00... P14141
1367930_at 29423 Gap43 Rat Expressi... Rattus norve... Rn.10928.1 Rn.10928 Nov. 2004 (B... growth asso... chr11q21 --- P07936
1367954_at 25454 Gfra1 Rat Expressi... Rattus norve... Rn.6281.1 Rn.88489 Nov. 2004 (B... glial cell line ... chr1q55 ENSRNOG00... O35748 /// Q...
1367961_at 24594 Ngfg Rat Expressi... Rattus norve... Rn.11331.1 Rn.11331 Nov. 2004 (B... nerve growt... chr1q22 ENSRNOG00... P00758 EC:3.4.2
1367970_at 81531 Pfn2 Rat Expressi... Rattus norve... Rn.3515.1 Rn.203100 Nov. 2004 (B... profilin 2 chr2q31 ENSRNOG00... Q9EPC6
1367973_at 24770 Ccl2 Rat Expressi... Rattus norve... Rn.4772.1 Rn.4772 Nov. 2004 (B... chemokine (... chr10q26 ENSRNOG00... P14844
1367986_at 29602 Ptgfrn Rat Expressi... Rattus norve... Rn.6332.1 Rn.6332 Nov. 2004 (B... prostaglandi... chr2q34 ENSRNOG00... Q62786
1367992_at 25719 Sgne1 Rat Expressi... Rattus norve... Rn.6173.1 Rn.6173 Nov. 2004 (B... secretory gr... chr3q34 ENSRNOG00... P27682
1368112_at 25539 Sag Rat Expressi... Rattus norve... Rn.9856.1 Rn.9856 Nov. 2004 (B... retinal S-anti... chr9q35 --- P15887
1368113_at 116592 Tff2 Rat Expressi... Rattus norve... Rn.34367.1 Rn.34367 Nov. 2004 (B... trefoil factor ... chr20p12 ENSRNOG00... Q09030
1368123_at 25718 Igf1r Rat Expressi... Rattus norve... Rn.10957.1 Rn.10957 Nov. 2004 (B... insulin-like gr... chr1q22 --- P24062 EC:2.7.1
1368172_a_at 24914 Lox Rat Expressi... Rattus norve... Rn.11372.1 Rn.11372 Nov. 2004 (B... lysyl oxidase chr18q11 --- P16636 EC:1.4.3
1368226_at 171047 RGD620382 Rat Expressi... Rattus norve... Rn.6997.1 Rn.6997 Nov. 2004 (B... Nucleoside 2... chr9q12 ENSRNOG00... O35820
1368243_at 29530 Amhr2 Rat Expressi... Rattus norve... Rn.10165.1 Rn.10165 Nov. 2004 (B... anti-Mullerian... chr7q36 ENSRNOG00... Q62893
1368269_at 25474 Lgals4 Rat Expressi... Rattus norve... Rn.9656.1 Rn.9656 Nov. 2004 (B... lectin, galact... chr1q21 ENSRNOG00... P38552
1368284_at 56765 Plvap Rat Expressi... Rattus norve... Rn.53915.1 Rn.53915 Nov. 2004 (B... plasmalemm... chr16p14 ENSRNOG00... Q9WVV78
1368303_at 63840 Per2 Rat Expressi... Rattus norve... Rn.25935.1 Rn.25935 Nov. 2004 (B... period homol... chr9q36 ENSRNOG00... Q9Z301

<Figure 3-10> Annotation Result Window

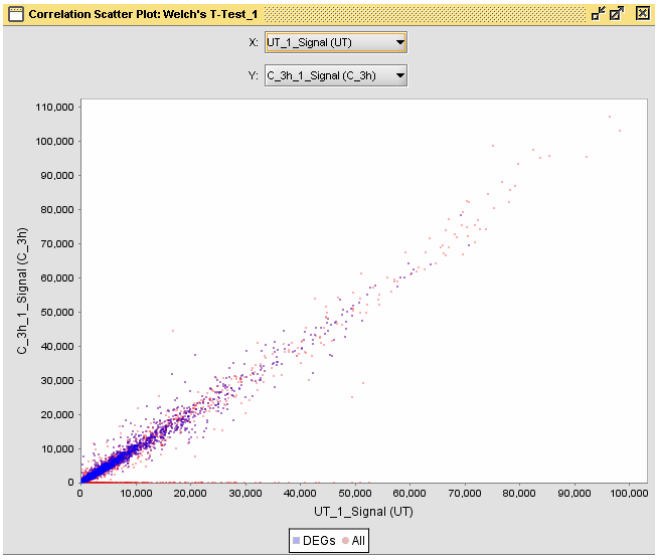
③ Clustering: Export to Clustering module only for selected genes.

④ Classification: Export to Classification module only for selected genes.

⑤ Pathway Analysis: Export to Pathway Analysis module only for selected genes.

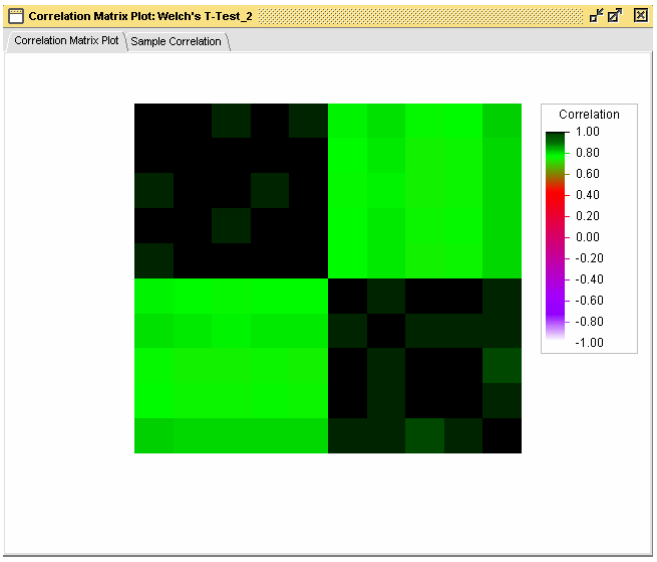
⑥ Plot

- Correlation Scatter Plot: This visual shows the relation between selected Differentially Expressed Genes



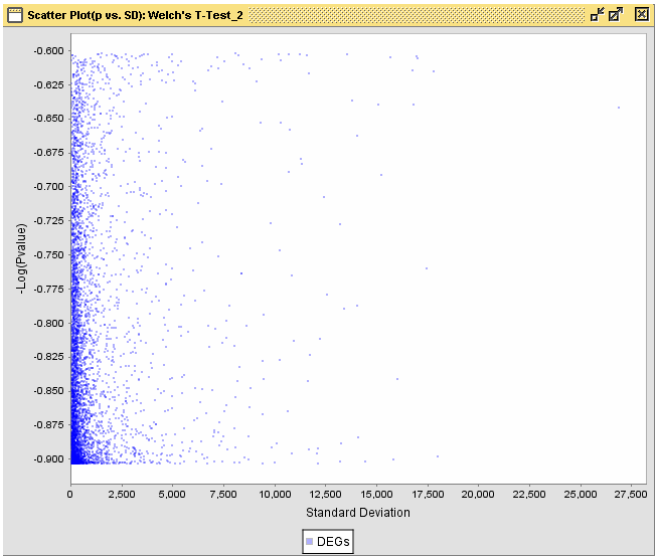
<Figure 3-11> Correlation Scatter Plot Result Window

- Correlation Matrix Plot: This visual shows the relation between selected Differentially Expressed Genes.



<Figure 3-12> Correlation Matrix Plot Result Window

- Scatter Plot (p vs. SD): This visual is taken with two axes, Standard Deviation and Log (P-value), on the Differentially Expressed Genes list in statistical method.

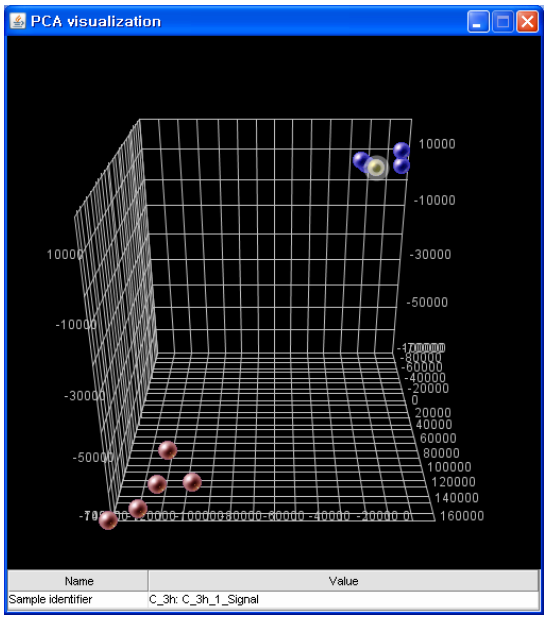


<Figure 3-13> Scatter Plot Result Window

- Sample PCA
  - Confirm the recurrence between samples expressed in 3D visual whether selected genes show the difference between groups, and can also confirm

the reliability of the Differentially Expressed Genes.

- Same groups are expressed in identical color. Click each ball to see the name of the sample on the Sample identifier.



<Figure 3-14> Sample PCA Result Window

⑦ DEG Filtering

- Minimum Signal Intensity: In the select Gene list, the user can confirm the result with genes deleted, which include smaller Signal value than the user have set up.
- Detection Call

  In case of One-Dye Chip Data, it is possible to confirm the result after applying Filtering based on Detection (PMA Call) of selected genes.

  - Remove the Probe IDs with Present Calls for less than □ arrays: Genes that Present Call is less than the number of Array in □ will be deleted.
  - Remove the genes with A or No Call grade at all arrays: Delete all genes which Detection is all A or No call in Array.
  - Select the genes with P grade at all arrays: You can confirm the result with genes which the average difference between signals of two classes is deleted from the selected gene list.
- Difference Between Averages (2-class only): You can confirm the result with genes which the average difference between signals of two classes is deleted from the selected gene list.

⑧ Browse tool and Heat Map Set Up

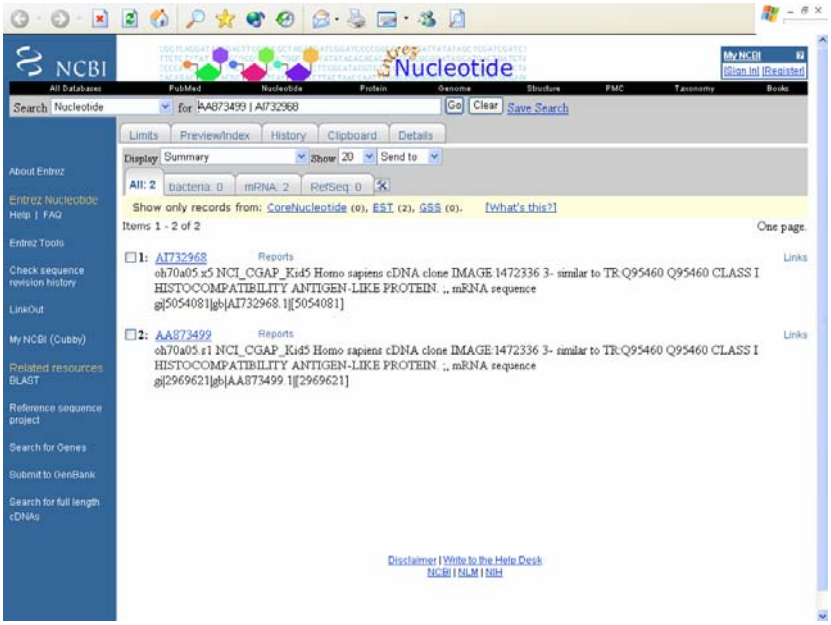
- Search: Search with ID or Line No. Select Search type, and input the ID or

Line No. on the text window, then click [Search] button to see the result table with reversed related gene. Then select [Case Sensitive] option to search separately, the capital letters and small letters.

- Image width: Change the width of Heat Map seen on the result window. Input the width to be adjusted and click [Enter] key.
- Select Heat Map: Change the color of Heat Map.
  - Red/Green
  - Blue/Yellow

⑨ List of DEG Finding result

- Line No: It shows the order of genes of input data.
- ID: It shows the ID of each gene. Click Hyper Link, then it will be linked to related database URL and it is possible to see the detailed information of corresponding genes. To have exact information, we must select exact ID Type when Analysis is creating.

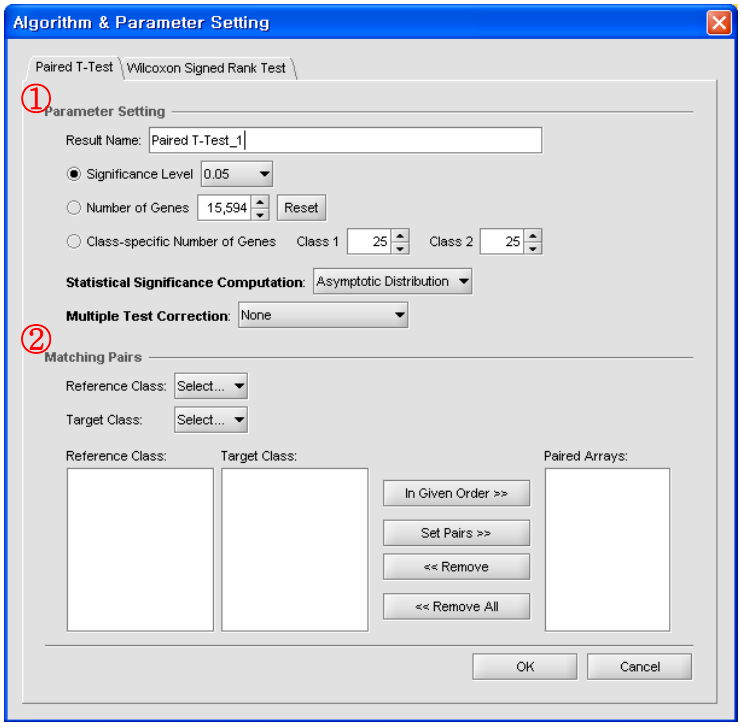


<Figure 3-15> Related Database URL Linked Window

- Average log2 (Fold Change): Shows Log2 (Fold Change) value of each gene.
- Average Fold change: Shows Fold Change value of each gene.
- Regulation: It is marked UP, when Average Log2 (fold change) value is bigger than cut off and marked DOWN, when smaller than cut off. This means that it is Up-regulated and Down-regulated.
- Heat Map: It shows the Heat Map of extracted expressed genes, and can

easily verify in one view, the intensity information of each gene. 3.3.2. 2-Class Paired Test.

3.3.2. 2-Class Paired Test



<Figure 3-16> Paired Test Parameter Set Up Window

3.3.2.1. Paired T-test

In the Paired T-Test [3-1], select [DEG finding] → [2-Class Paired Test] → [Paired T-test] on the menu bar, then the Parameter set up window appears.

① Parameter Setting

- Significance Level: Select significant genes below level that user has set up.
- Number of Genes: Select genes in higher scored order that user have set up.
- Class specific Number of Genes: Select genes in higher score order of class.
- Statistical Significance Computation: Select the method to seek for P-value.
  - Asymptotic Distribution: In case of assuming the data ratio distribution as the regular distribution
  - Permutation Test: In case of no assumption of data ratio distribution

- Multiple Test Correlation: Select the method to revise P-value
  - None
  - Bonferroni
  - Holm's procedure
  - Benjamini-Hochberg FDR

② Matching Pairs

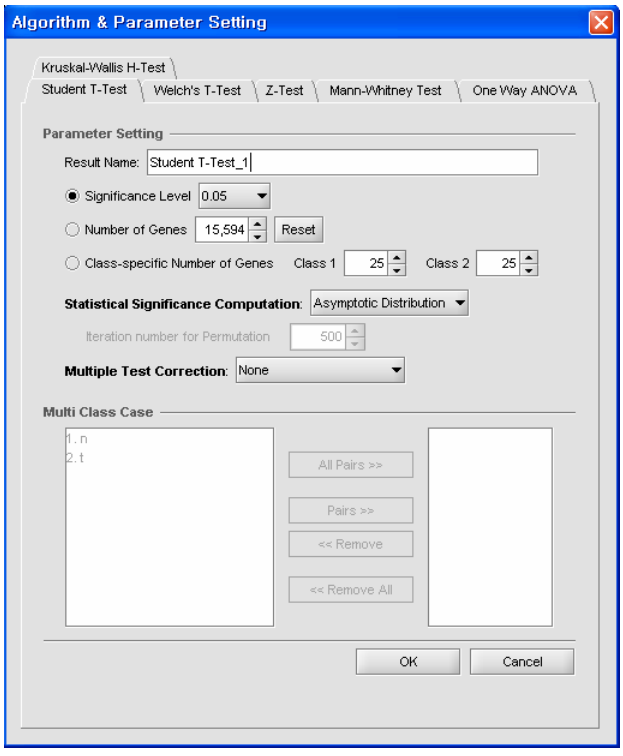
- Select Reference class and Target class, and then set up the sample Pair.
- [In Given Order>>]: Set up the pair with given order.
- [Set Pairs>>]: Set up the sample pair which user have selected.
- [<<Remove], [<<Remove All]: Set up or cancel pair.

③ Click [OK] button to see the result.

3.3.2.2. Wilcoxon Signed Rank Test

In Wilcoxon Signed Rank Test[3-1], select [DEG finding] → [2-Class Paired Test] → [Wilcoxon Signed Rank Test] on the menu bar, then set up window appears. Parameter is identical with Paired T-Test (Ref.: 3.3.2.1)

3.3.3. 2-Class Unpaired Test



<Figure 3-17> 2-Class Unpaired Test Parameter Set Up Window

3.3.3.1. Student T-test

In Student T-Test, just select [DEG finding] → [2-Class Unpaired Test] → [Student T-Test] on the menu bar. Parameter is identical with Welch's T-test (Ref.: 3.3.3.2)

3.3.3.2. Welch's T-test

In Welch's T-Test[3-2], select [DEG finding] →[2-Class Unpaired Test] → [Welch's T-test] on the menu bar, then set up window appears.

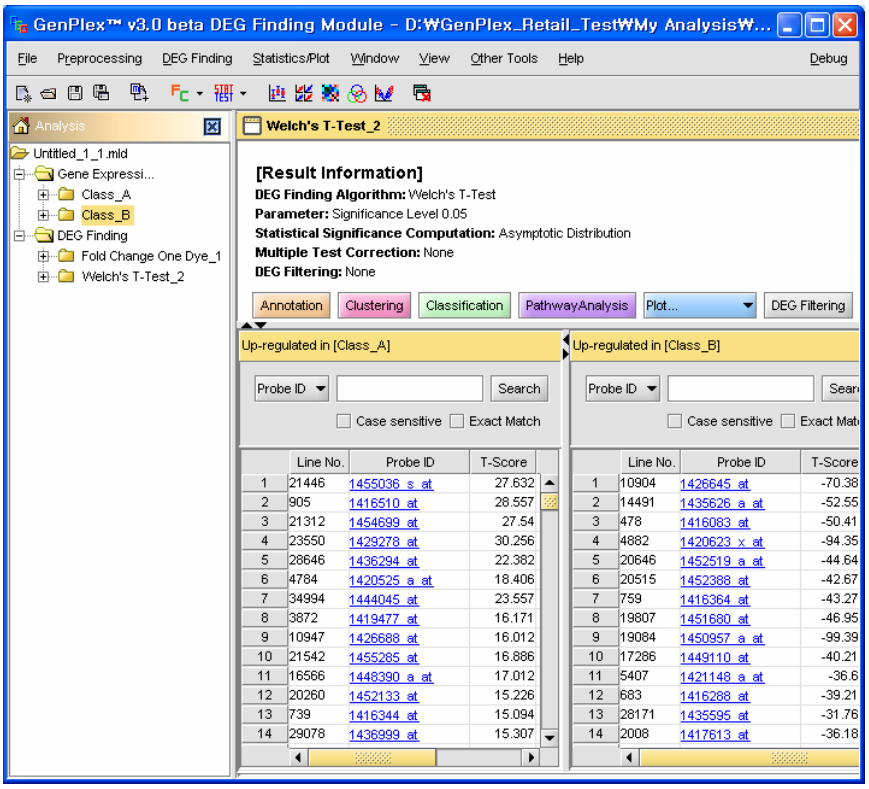
① Parameter Setting

- Significance Level: Select genes below significant standard that user have set up.
- Number of Genes: Select genes in higher score order that user have set up the number.
- Class specific Number of Genes: Select genes in higher score order of class.
- Statistical Significance Computation: Select the method of P-value.
  - Asymptotic Distribution: In case of assuming the data ratio distribution as regular distribution
  - Permutation Test: In case of not assuming the data ratio distribution
- Multiple Test Correlation: Select the method to revise P-value
  - None
  - Bonferroni
  - Holm's procedure
  - Benjamini-Hochberg FDR

② Multiple Class Case

If the Class is more than two, it is possible to select two classes and apply Welchh's T-Test. Select two classes from the list on the left side of the window (use Ctrl or Shift key), then click [pairs>>] button to set up.

③ Click [OK] button to see the result window (Ref.: 3.3.1.3)



<Figure 3-18> Welch's T-test Result Window

3.3.3.3. Z-test

In the Z-Test[3-3], just select [DEG Finding] → [2-Class Unpaired Test] → [Z-Test] from the menu bar. Parameter is identical with Welch's T-Test (Ref.: 3.3.3.2)

3.3.3.4. Mann-Whitney Test

In Mann-Whitney Test[3-1], just select [DEG finding] → [2-Class Unpaired Test] → [Mann-Whitney Test] on the menu bar. Parameter is identical with Welch's T-Test (Ref.: 3.3.3.2)

3.3.4. Multi-Class Test

This is used in comparing more than 3 classes statistically.

3.3.4.1. One Way ANOVA

In One Way ANOVA[3-1], just select [DEG Finding] → [Multi-Class Test] → [One Way ANOVA] on the menu bar. Parameter is similar to Welch's T-test (Ref.: 3.3.3.2)

3.3.4.2. Kruskal-Wallis H-test

In Kruskal-Wallis H-Test[3-1], just select [DEG Finding] → [Multi-Class Test] → [Kruskal-Wallis H-Test] on the menu bar. Parameter is similar to Welch's T-Test (Ref.: 3.3.3.2)

3.3.5. Combine Results

It is possible to combine genes of DEG Finding Results in various methods. Select [DEG Finding] → [Combine Results] on the menu bar, then the set up window appears.

Select DEG Finding Results to be combined from the list, then click [Combine] button to see the result. [AND], [OR], [Complement] operation is possible to figure out the Intersection, Union and Complement of each list. For [Common Gene Count], it will show gene list common in more than the user have set up from the result selected from the list.

3.3.6. Import Gene List

In case of the text file in form of gene ID input in each row, it is possible to input this for the result of DEG Finding. On the menu bar, select [DEG Finding] →[Import Gene List], then the Data Selecting Window appears.

3.3.7. Export to Clustering Module

On the menu bar, select [DEG Finding] → [Export to Clustering Module], it is possible to export various DEG Finding Results in to Clustering module.

3.3.8. Export to Pathway Analysis Module

On the menu bar, select [DEG Finding] → [Export to Pathway Analysis Module], it is possible to export various DEG Finding Results in to Pathway Analysis module.

3.3.9. Save Result(s) As Text

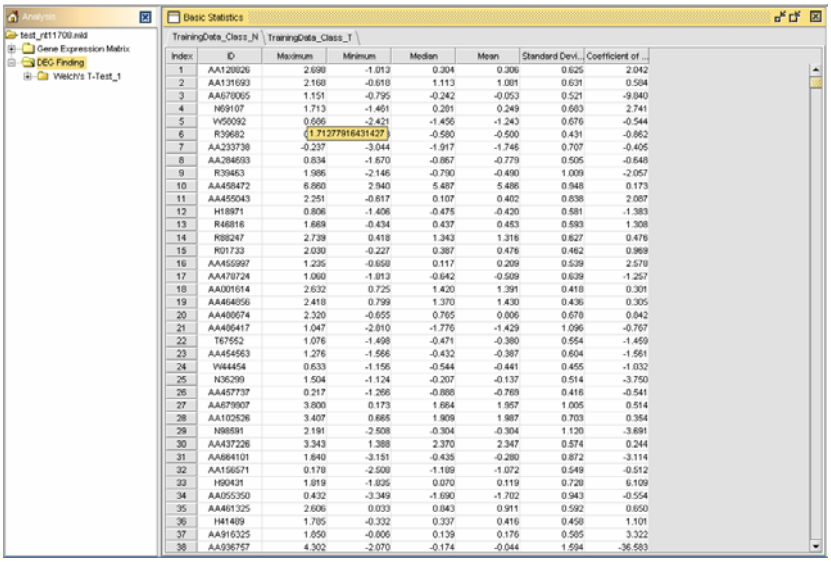
On the menu bar, select [DEG finding] → [Save Result(s) As Text...], it is possible to select DEG Finding Result and save it in to text file.

3.4. Statistics/Plot

It is possible to verify the input data statistically through various plots.

3.4.1. Basic Statistics

You can verify basic statistics of the input data. On the menu bar, select [Statistics/Plot] → [Basic Statistics], then user can see the result.



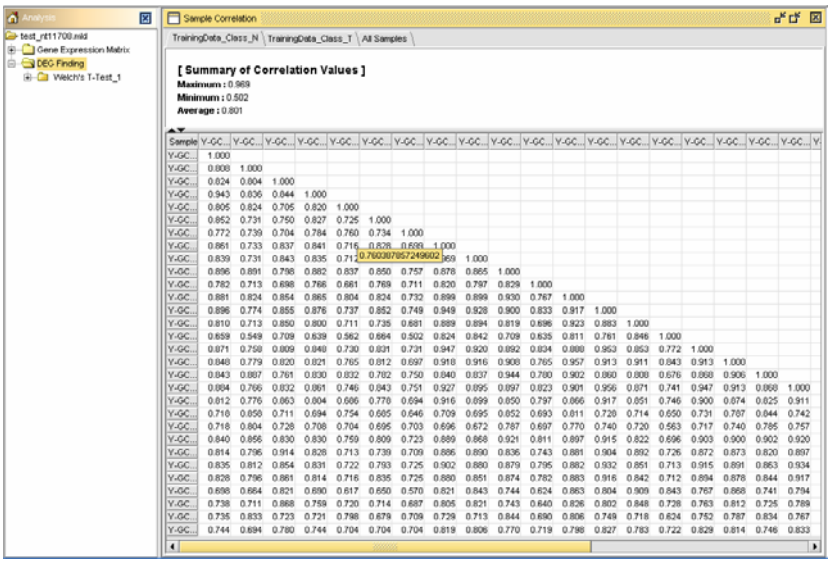
<Figure 3-19> Basic Statistics Result Window

① The result of each Class is shown separately in tab.

- ID: ID of gene
- Maximum: Greatest value of classified gene
- Minimum: Smallest value of classified gene
- Median: Median value of classified gene
- Mean: Average (Mean) of classified gene
- Standard Deviation: Standard Deviation of classified gene
- Coefficient of Variation: CV value of classified gene

3.4.2. Sample Correlation Matrix

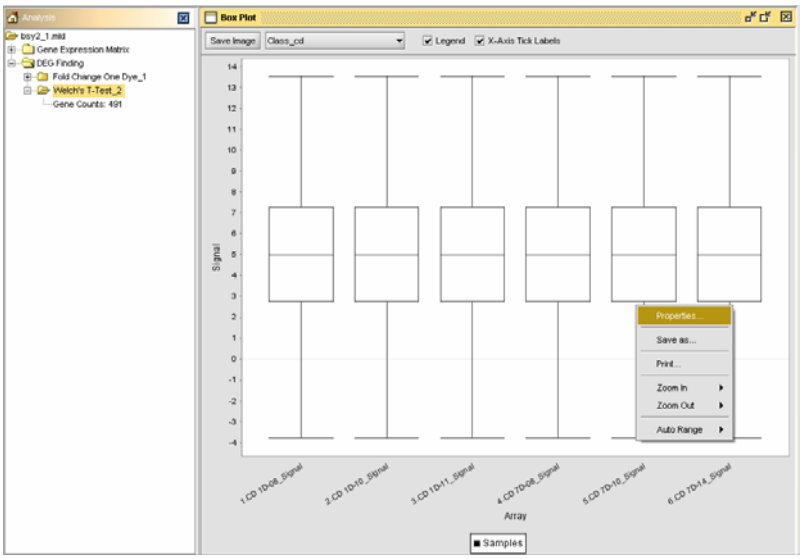
This shows the Correlation of the Sample. On the menu bar, select [Statistics/Plot] → [Sample Correlation], then user can verify the result. Each Class Result is shown separately in the tab.



<Figure 3-20> Sample Correlation Matrix Result Window

3.4.3. Box Plot

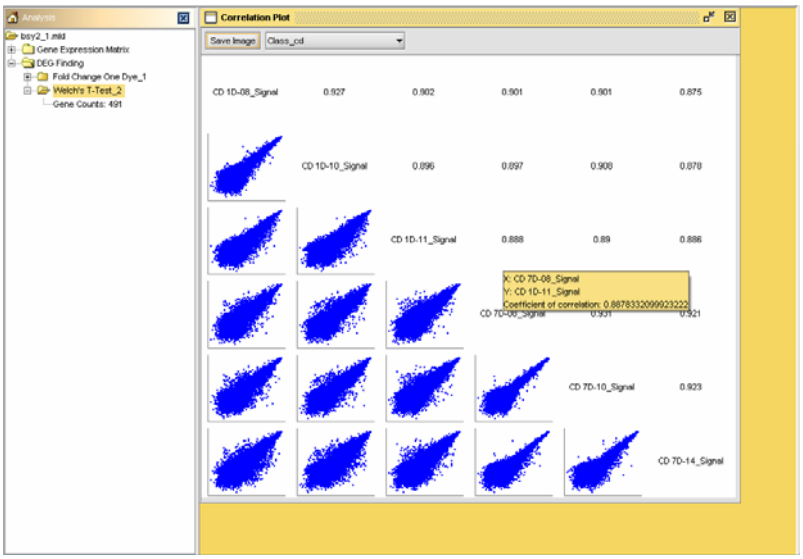
On the menu bar, select [Statistics/Plot] → [Box Plot], or click on the eighth icon , to see the Box Plot for each Class.



<Figure 3-21> Box Plot Result Window

3.4.4. Correlation Scatter Plot

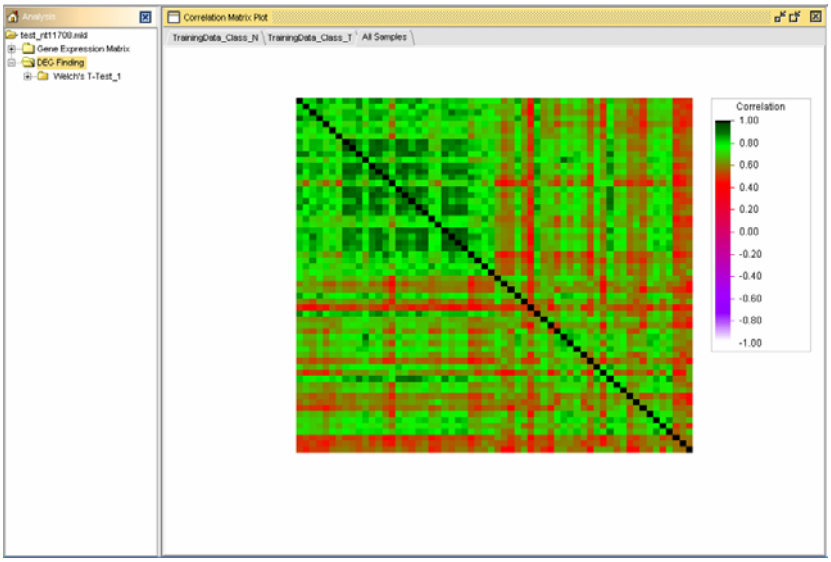
On the menu bar, select [Statistics/Plot] → [Correlation plot], or click on the ninth icon , to see the Correlation Plot for each Class.



<Figure 3-22> Correlation Scatter Plot Result Window

3.4.5. Correlation Matrix Plot

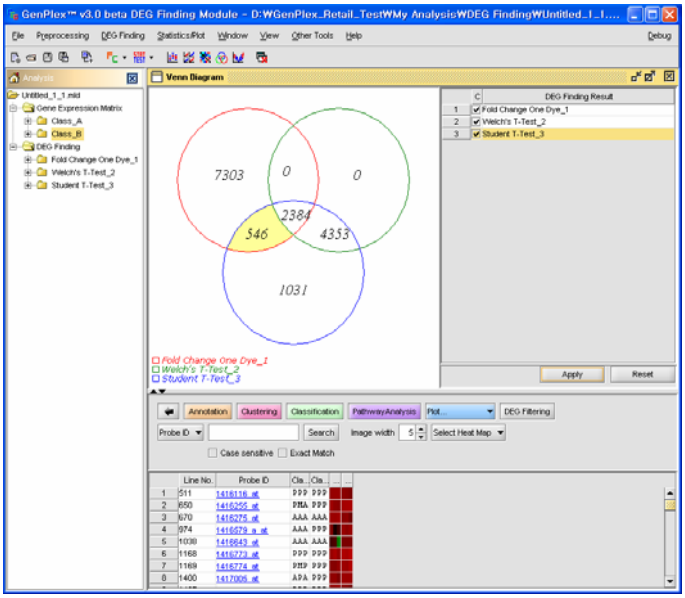
On the menu bar, select [Statistics/Plot] → [Correlation Matrix Plot], or click on the tenth icon , to see Correlation Matrix Plot for each Class.



<Figure 3-23> Correlation Matrix Plot Result Window

3.4.6. Venn Diagram

On the menu bar, select [Statistics/Plot] → [Venn Diagram], or just click on the eleventh icon , we can verify the Venn Diagram and gene list of each combination with 2~3 DEG Finding Result.

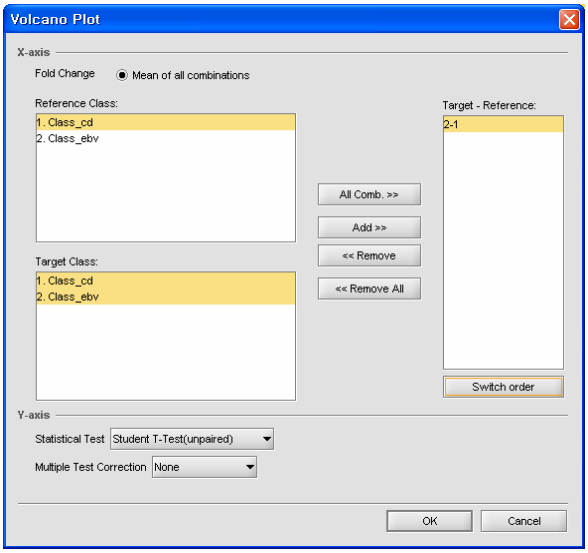


<Figure 3-24> Venn Diagram Result Window

① Select 2 or 3 DEG Finding Results from the table on the right side of the window and click [Apply] button, then the result will be shown on the Venn Diagram on the left side. Click [ ] button on the result below the table, corresponding gene list will be added in the tree on the left side as DEG Finding result.

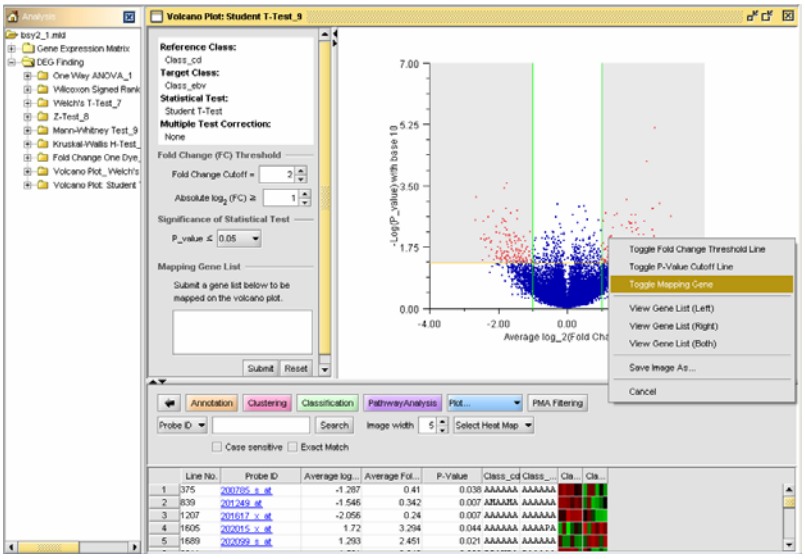
3.4.7. Volcano Plot

In Volcano Plot[3-4], select [Statistics/Plot] → [Volcano Plot] on the menu bar, or just click on the twelfth icon .



<Figure 3-25> Volcano Plot Set Up Window

① Set up Reference and Target Class to add in the list on the right side of the window, select Statistical Test, then click [OK] button to see the Volcano Plot Result Window.



<Figure 3-26> Volcano Plot Result Window

② Control Fold Change Threshold and P-value to have them reflected on the right side Plot. It is possible to verify the distribution in Plot adding, already known gene list or wanted gene list.

③ Double click on the wanted range on the right side Plot, to verify the information of corresponding gene of the selected range. Click [⬅] button to add the corresponding gene list as DEG Finding result on the gene list on the left side of the window.

3.5. Reference

[3-1] J.H. Zar, 'Biostatistical Analysis', 4th Edition, Prentice Hall Inc.

[3-2] B.L. Welch (1947) The generalization of 'students' problem when several different population variances are involved. Biometrika, 34:28-35.

[3-3] J.G. Thomas, J.M. Olson, S.J. Tapscott, L.P. Zhao (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res. 11:1227-1236.

[3-4] X. Cui & G.A. Churchill (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 4:210.

Clustering

4. Clustering

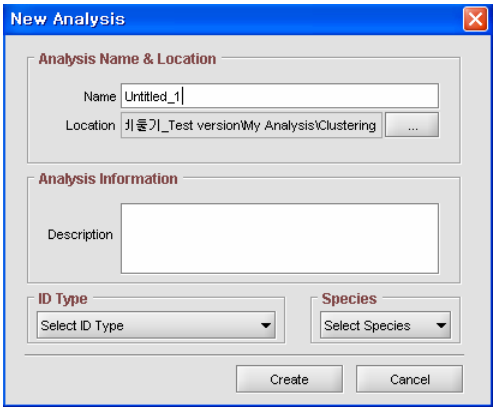
This is the method of gene Clustering or sample Clustering according to similar significant pattern. Gene Clustering is used for gene function search, and Sample Clustering is mostly studied for diagnosis, prognosis and prediction in medical field. Clustering module provides various clustering method and visualization. Also it is possible to operate statistical verification of Clustering result.

4.1. File

4.1.1. New Analysis

If data is exported from the Preprocessing module, Analysis File is automatically created. So this procedure can be skipped (go to 4.2 ▷).

On the menu bar, select [File] → [New Analysis], or click on the first icon , then Analysis creating window appears.



<Figure 4-1> Analysis Creating Window

① Name: Input the name of the Analysis File to be created.
② Location: Click [⋯] button to select location where Analysis File will be created.
③ Description: Input supplement information in to Analysis File (can be skipped).
④ Probe ID Type: Select ID Type of the input data.

- Commercial Product Probe ID
  - Affymetrix GeneChip Probe ID
  - Agilent Probe ID(One-dye)
  - Agilent Probe ID(Two-dye)
  - Applied Biosystems 1700 Probe ID
  - CodeLink Probe ID

- Illumina Probe ID
- Operon Probe ID

- Public DataBase ID
  - IMAGE Clone ID
  - NCBI Clone ID
  - NCBI GenBank Accession
  - NCBI GeneID (LocusLink)
  - NCBI UniGene ID
- Others (If ID is unknown)

⑤ Species: Select the species of the input data.

⑥ [Click [Create] button, then Analysis File will be created and data selecting window will appear (go to 4.1.7 ▷).

4.1.2. Open Analysis

On the menu bar, select [File] → [Open Analysis], or click on the second icon to open the saved Analysis File.

4.1.3. Recent Analysis

On the menu bar, select [File] → [Recent Analysis] to open updated analyzed Analysis File. This list can be deleted using [Clear History] menu.

4.1.4. Save Analysis

On the menu bar, select [File] → [Save Analysis], or click on the third icon to save working Analysis File.

4.1.5. Save Analysis As

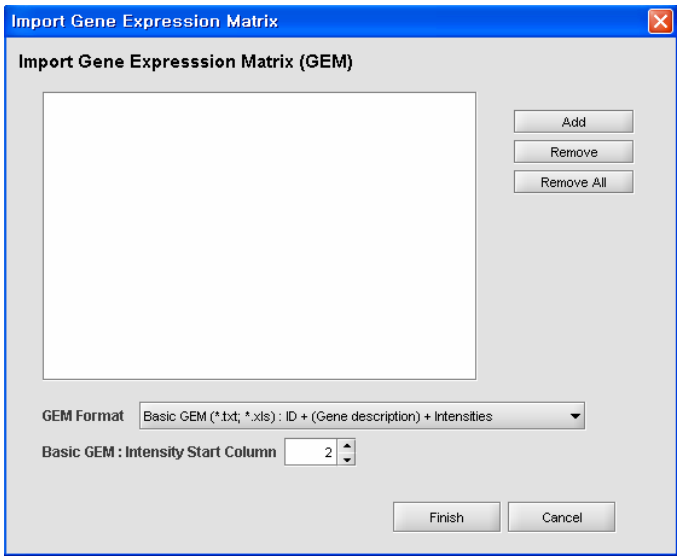
On the menu bar, select [File] → [Save Analysis As...], or click on the fourth icon to save working Analysis File in different name.

4.1.6. Close Analysis

On the menu bar, select [File] → [Close Analysis] to close working Analysis File.

4.1.7. Import Data

On the menu bar, select [File] → [Import Data], then the data selecting window appears.



<Figure 4-2> Data Selecting Window

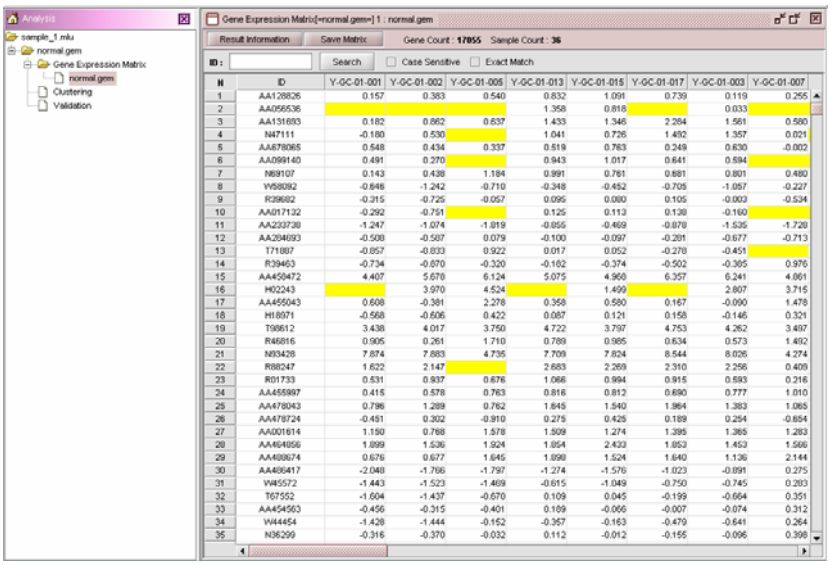
① Click [Add] button to select input file, then it will be added in the list on the left side of the window. Use [Remove] and [Remove All] button to delete item.

② GEM Format

- Basic GEM: ID+ (Gene description)+ Intensities
  - Intensity Start column: In case of Basic GEM, it is possible to fix the location where column is started. It is also possible to use description information in case there is description column between ID column and intensity.
- Basic GCOS output: Signal+ Detection+ Detection p-value
- Basic GCOS output: Signal+ Detection

③ Click [Finish] button to input data.

■ Input Data Result



<Figure 4-3> Input Data Result Window

① To see the input data, double click the name of the input data in the search window. Missing value is marked in yellow.

4.1.8. Analysis Properties

On the menu bar, select [File] → [Analysis Properties] to adjust information of working Analysis File.

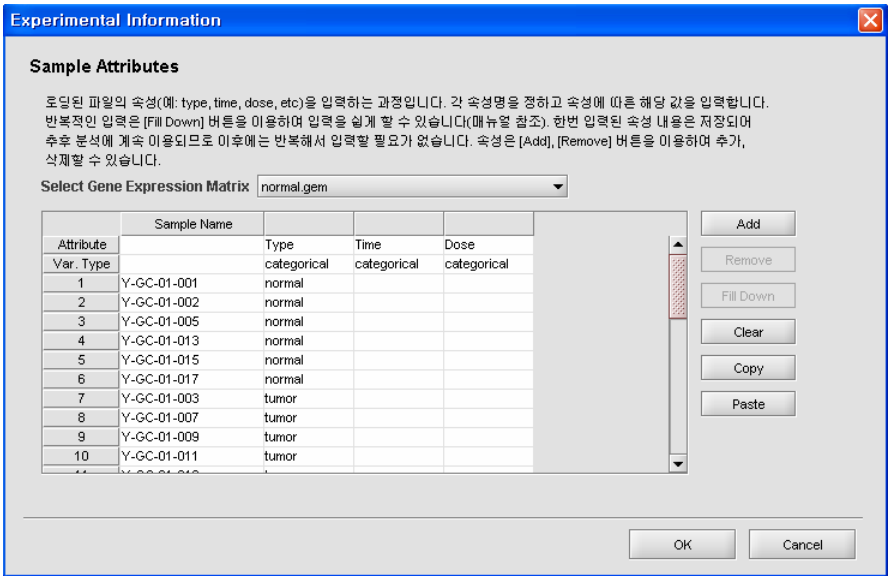
4.1.9. Exit

On the menu bar, select [File] → [Exit] to exit the program.

4.2. Preprocessing

4.2.1. Experimental Information

This is the procedure to input the experiment information of the input data, just select [Preprocessing] → [Experimental Information] on the menu bar.



<Figure 4-4> Experimental Information Input Window

① Select Gene Expression Matrix: Select GEM to input or change the attribute.

② Attr. Name: Input the name of the attribute. Type, Time, Dose are input as basic attribute. It is possible to adjust these attributes, and use [Add], [Remove] button to add or delete attributes.

③ Var. Type: This is the type of attribute, can select Categorical or Continuous (Now supporting Categorical type only).

④ Double click the cell to input the attribute, use [Fill Down], [Copy], [Paste] button for easier input.

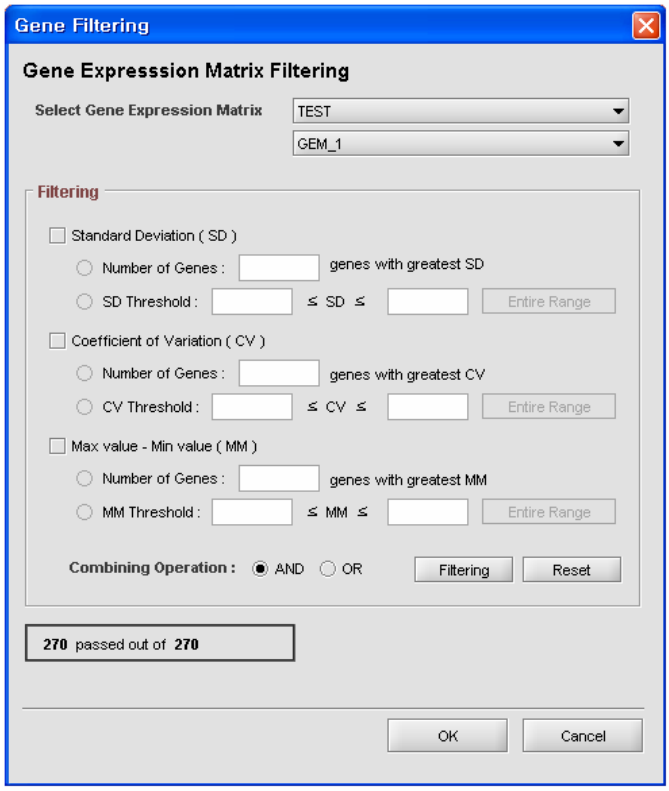
⑤ Click [OK] button, the input attribute will be applied.

4.2.2. Log Transform

On the menu bar, select [Preprocessing] → [Log Transform] to transform input data into Log value.

4.2.3. Gene Filtering

On the menu bar, select [Preprocessing] → [Gene Filtering], or click on the seventh icon to select high ranking genes using statistical method of input data.



<Figure 4-5> Gene Filtering Set Up Window

① Select Gene Expression Matrix: Select GEM which Filtering will be applied.

② Filtering Option

- Standard Deviation (SD): Select genes same as the figure that the user has set up in bigger SD order.
- Coefficient of Variation (CV): Select genes same as the figure that the user has set up in bigger CV order.
- Max value – Min Value (MM): Select genes same as the figure that the user have set up in bigger difference between Maximum Value and Minimum Value.

③ Click [Filtering] button, selected gene figure are shown below.

④ [Click [OK] button to add filtered GEM.

4.2.4. Missing Data Filtering

On the menu bar, select [Preprocessing] → [Missing Data Filtering] to delete missing entry optionally from input data (Ref.: 3.2.2)

4.2.5. Imputation

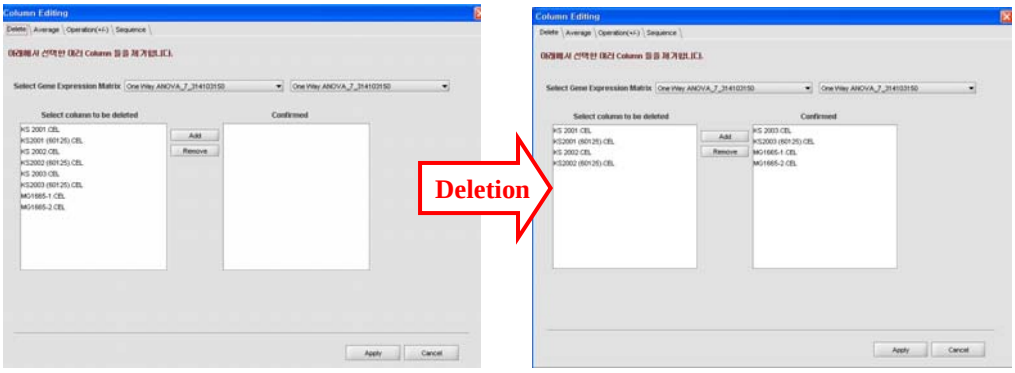
On the menu bar, select [Preprocessing] → [Imputation] to fill up missing value according to regular rule.

4.2.6. Column Editing

On the menu bar, select [Preprocessing] → [Column Editing] to edit or combine each sample of input data in various ways.

① Delete

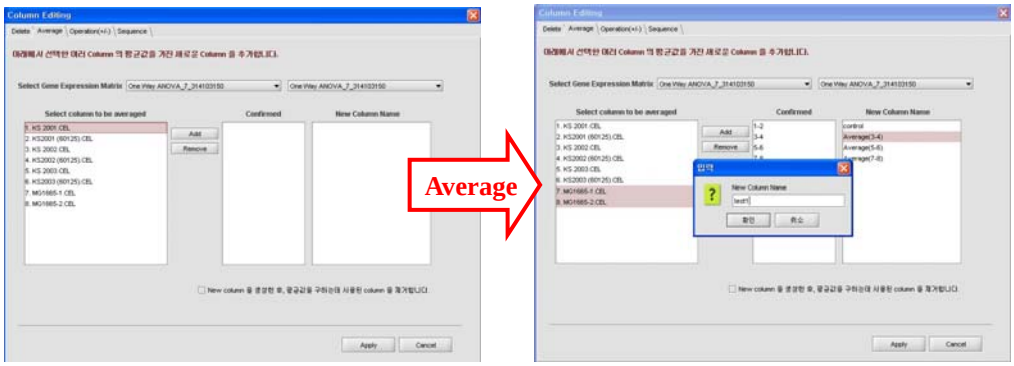
It can create GEM excluding selected sample.



<Figure 4-6> Column Editing: Delete Set Up Window

② Average

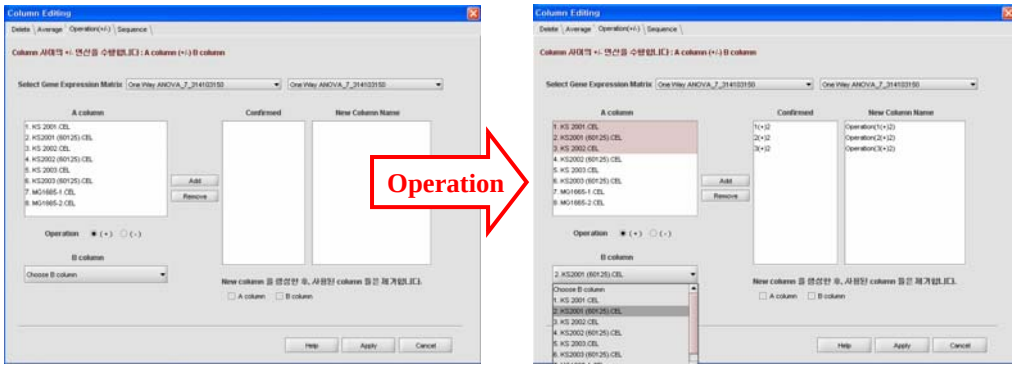
Create GEM adding column with average value of selected sample. Double click on each item of the new column to change the name.



<Figure 4-7> Column Editing: Average Set Up Window

③ Operation (+/-)

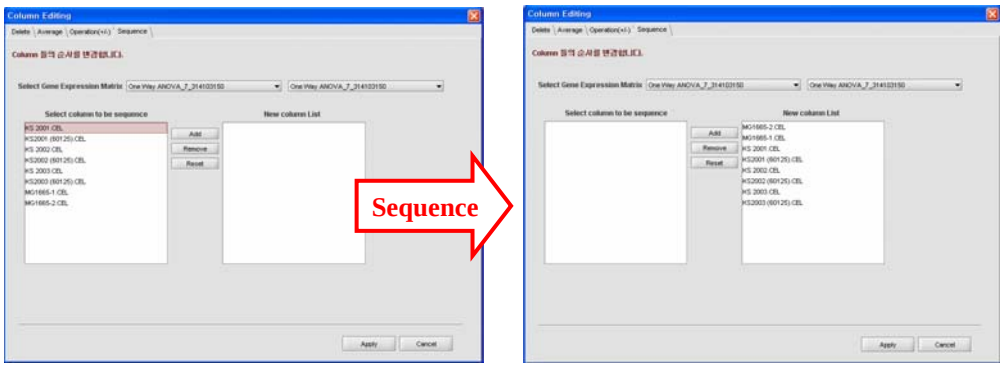
Create GEM adding column with value added or extracted of reference sample of selected sample. Double click on each item of the new column to change the name.



<Figure 4-8> Column Editing: Operation (+/-) Set Up Window

④ Sequence

Create GEM with column order relocated.



<Figure 4-9> Column Editing: Sequence Set Up Window

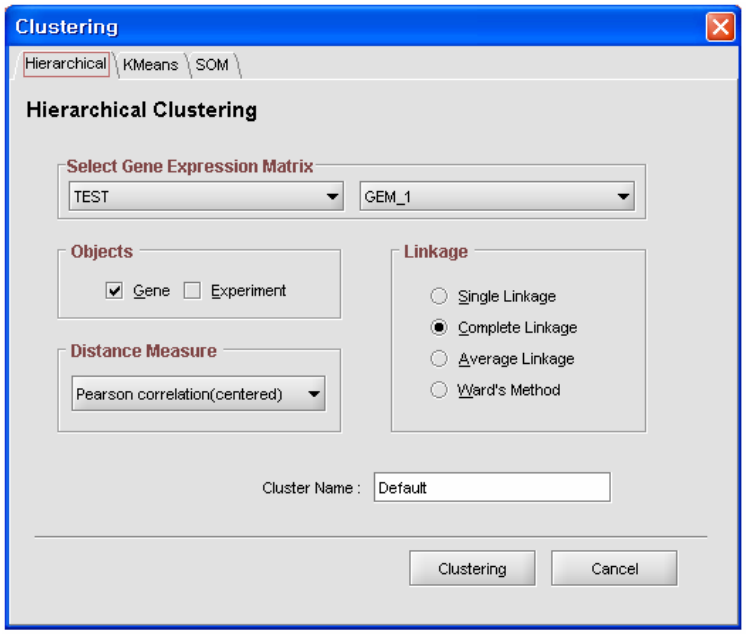
⑤ Rename

Change the name of selected sample.

4.3. Clustering

4.3.1. Hierarchical Clustering

For Hierarchical Clustering[4-1], select [Clustering] → [Hierarchical Clustering]on the menu bar, or click on the eighth icon to see set up window.



<Figure 4-10> Hierarchical Clustering Set Up Window

① Select Gene Expression Matrix: Select GEM when apply Clustering.

② Objects: Select standard to apply Clustering.

- Gene: Standard on gene.
- Experiment: Standard on Sample.

③ Distance Measure: This is used to calculate the distance between two individuals (Ref.: 8.1).

- Euclidean Distance: Geometrical distance between two individuals
- Manhattan Distance: Distance between two individuals considering the importance of each variables
- Pearson Correlation(centered): Measure similarity of two individuals using coefficient of correlation after transforming the average of each individuals in 0, and decentralization to 1

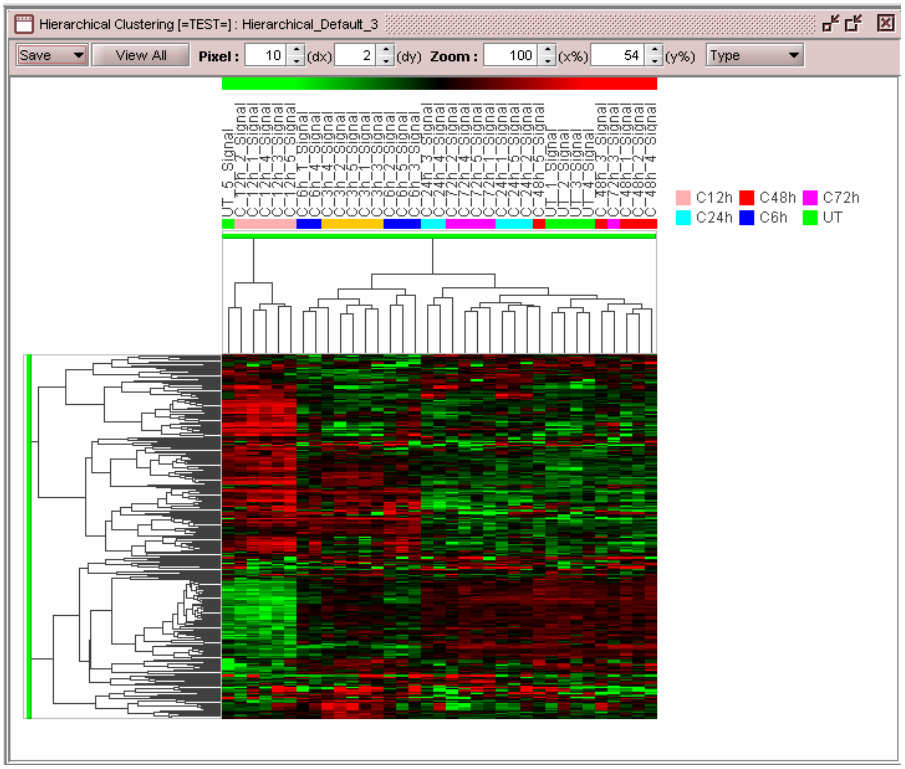
- Pearson Correlation (uncentered): Measure similarity of two individuals using calculated correlation coefficient with actual signal value of two individuals.
- Absolute Pearson: Use the absolute value of Pearson correlation coefficient

④ Linkage: This is the method to calculate the distance between Clusters.

- Average Linkage: This is the method to adjust in to similarity of entire Clusters after having the outcome of the similarity average between all individuals in two Clusters composing a new Cluster.
- Complete Linkage: This is the method to adjust the lowest similarity value in to similarity value of entire Clusters, among the similarity value between all individuals in two Clusters composing a new Cluster.
- Single Linkage: This is the method to adjust the highest similarity value in to similarity of entire Clusters, among the similarity value between all individuals in two Clusters composing a new Cluster.
- Ward's Method: This is the method to operate clustering in a way of minimizing after calculating the sum of the squares among the group, from the average value of each cluster to each individual after calculating the average value of each cluster on all variables.

⑤ Input the name of the Clustering Result, and click [Clustering] button to verify the Clustering Result (Dendrogram).

■ Hierarchical Clustering Result (Dendrogram)



<Figure 4-11> Hierarchical Clustering Result Window (Dendrogram)

① Click [Matrix] button to save GEM in to text file. Click [Image] button to save Dendrogram in to picture file. Click [Initialize] button to adjust the cell size of the Dendrogram using basic set up size or fix whole screen. Click [X] and [Y] button to control the width and length of the cell size.

■ Dendrogram Pop Up Menu

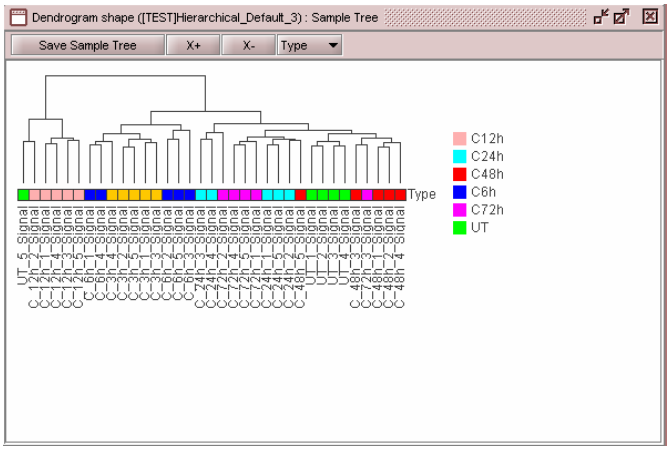
Right click on the mouse in Dendrogram to see the pop-up menu.

① Heatmap Color: Scale around row average: Control the color of Up/Down, based on each average of gene.

② Heatmap Color: Yellow/Blue (up/down): Can change the color of Heat Map in to Yellow and Blue.

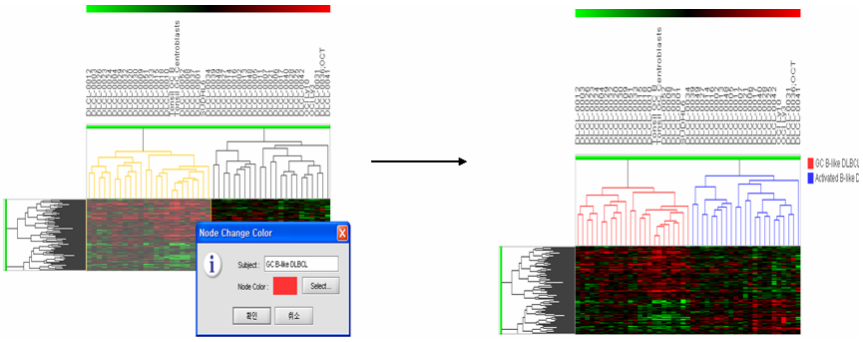
③ Heatmap Color: Brightness Scale: Can control the brightness of Heat Map.

④ Dendrogram Shape: Sample Tree: It only shows the Sample Tree of the Dendrogram.



<Figure 4-12> Dendrogram Shape: Sample Tree Result Window

⑤ Dendrogram Branch Coloring: After selecting each Node from Dendrogram, click this menu, then it is possible to fix the name and color of Node.



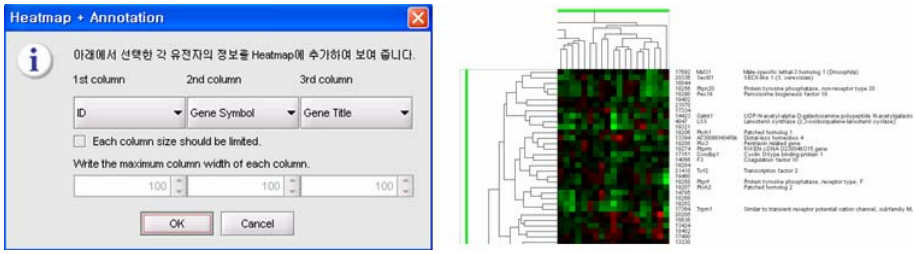
<Figure 4-13> Select Node Change Color Set Up Window (Left) and Set Up Result (Right)

⑥ Reset Branch Coloring: Can reset the color of Node.

⑦ Dendrogram Color scale bar: Can see the color of scale bar.

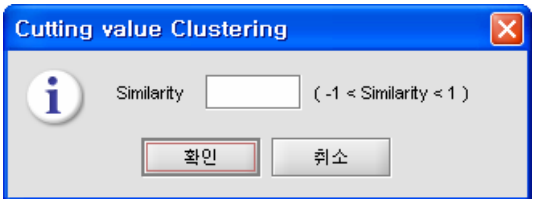
⑧ Retrieve Annotation Data: Can verify the Annotation Information of genes corresponding with the selected Node.

⑨ Heatmap+ Annotation: Can see in one view, the Annotation Information on the right side of the Dendrogram. But not Illumina Probe ID.



<Figure 4-14> <Figure 4-13> Select Node Change Color Set Up Window (Left) and Set Up Result (Right)

⑩ Branch-cut Value: Can divide the cluster inputting Distance Measure that user has input.



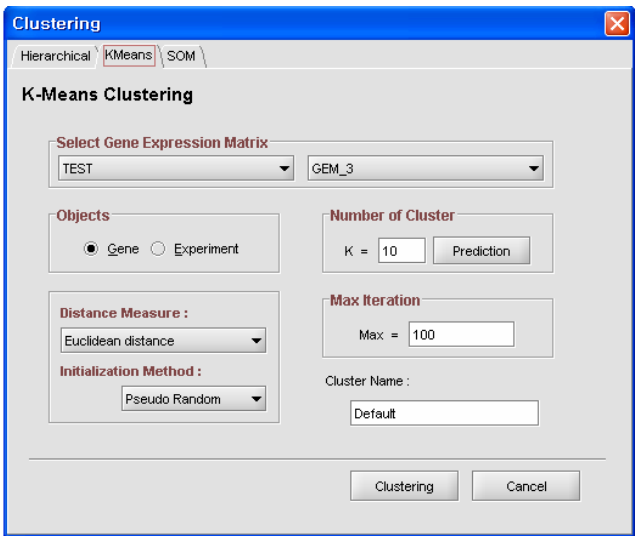
<Figure 4-15> Cutting value Clustering Set Up Window

⑪ Create Cluster: In case of fixing the Cluster that user have input moving the green Moving Bar, or input the Branch-cut Value. Based on this, user can create each Cluster and verify the result.

⑫ Save Sub Tree Matrix: Can save GEM data of Node selected from Dendrogram in to text file.

4.3.2. K-means Clustering

In K-means Clustering [4-2] [4-3], select [Clustering] → [K-means Clustering] on the menu bar, or click on the ninth icon , then set up window appears.



<Figure 4-16> K-means Clustering Set Up Window

① Select Gene Expression Matrix: Select GEM to apply Clustering.

② Objects: Select standard to apply Clustering.

- Gene: Standardize the gene.
- Experiment: Standardize the Sample.

③ Distance Measure: Select the method used for the calculation of Clustering distance (Ref.: 8.1).

- Euclidean Distance: Geometrical distance between two individuals
- Manhattan Distance: 각 Distance between two individuals considering specific gravity that each variable occupies
- Pearson Correlation(centered): Measure the similarity grade of two individuals using the correlation coefficient after transforming each individual's average 0, and diversity 1
- Pearson Correlation (uncentered): Measure the similarity grade of two individuals using actual signal value calculated correlation coefficient of two individuals.
- Absolute Pearson: Use the absolute value of Pearson correlation coefficient.

④ Initialization Method: Select the method of initialization.

- Pseudo Random: Generate similar random number every repetition
- Totally Random: Generate random number optionally every repetition

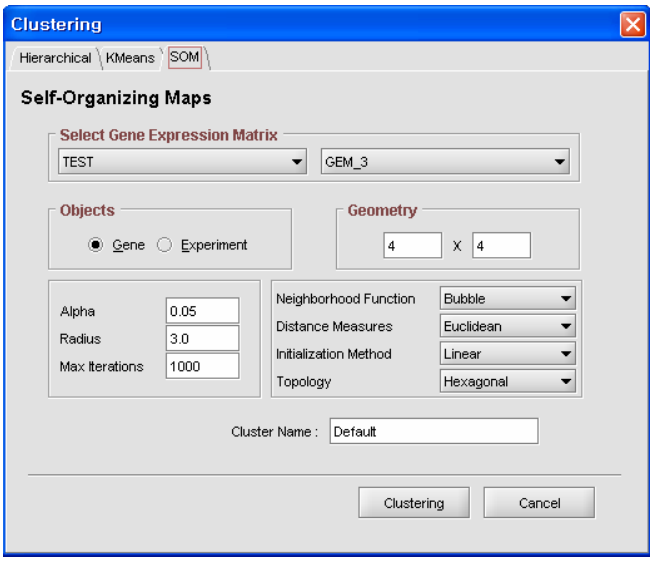
⑤ Number of Cluster: Input the number of Cluster. Click [Prediction] button to search for the most suitable K value first, then continue the operation (Ref.: 4.4.2).

⑥ Max Iteration: Input maximum repetition frequency (Basic value: 100).

⑦ Input the name of the Clustering Result, then click [Clustering] button to verify the Clustering Result.

4.3.3. Self Organizing Map

In Self Organizing Map (SOM) [4-4], select [Clustering] → [Self Organizing Map] on the menu bar, or click on the tenth icon , then the set up window appears.



<Figure 4-17> Self Organizing Map Set Up Window

① Select Gene Expression Matrix: Select GEM which will be applying Clustering.

② Objects: Select the standard applied with Clustering.

- Gene: Standardize the gene.
- Experiment: Standardize the Sample.

③ Geometry: Fix the number of Cluster in second dimension Geometry form (Basic value: 4×4).

④ Possible to fix Initial Alpha Value (Basic value: 0.05), Radius Value (Basic value: 3.0), Max. Iteration Value (Basic value: 1,000).

⑤ Select Mathematical function composing SOM.

- Neighborhood Function
  - Bubble
  - Gaussian
- Distance Measure (Ref.: 8.1)
  - Euclidean Distance : Geometrical distance between two individuals
  - Manhattan Distance: Distance between two individuals considering the

gravity occupying in each variable

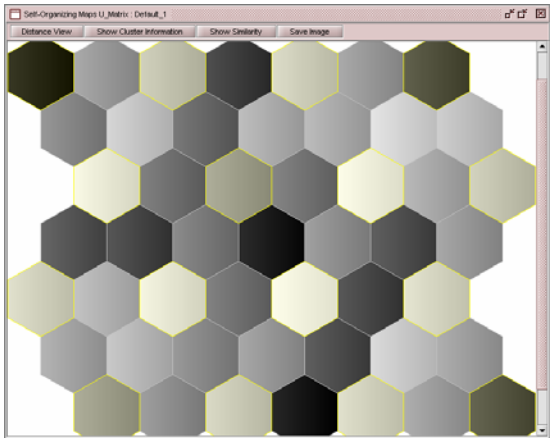
- Pearson Correlation(centered): Measures the similarity grade of two individuals using the correlation coefficient after transforming each individual's average 0, and diversity 1
- Pearson Correlation (uncentered): Measures the similarity grade of two individuals using actual signal value calculated correlation coefficient of two individuals
- Absolute Pearson: Use the absolute value of Pearson correlation coefficient

- Initializing Method
  - Linear
  - Random
- Topology
  - Hexagonal: Define the Neighborhood radius in hexagon form
  - Rectangular: Define the Neighborhood radius in rectangular form

⑥ Input the name of the Clustering Result and click [Clustering] button to verify Clustering Result.

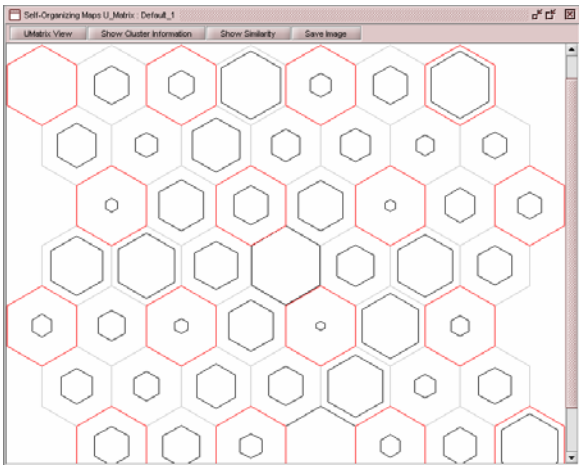
■ Self Organizing Map Result

It is easy to classify the similarity between each Cluster with color, and provide various options if the user click the button on the top of the Result Window.



<Figure 4-18> Self Organizing Map Result Window (U-Matrix)

- Distance View: As seen below figure, it is easy to classify the similarity.



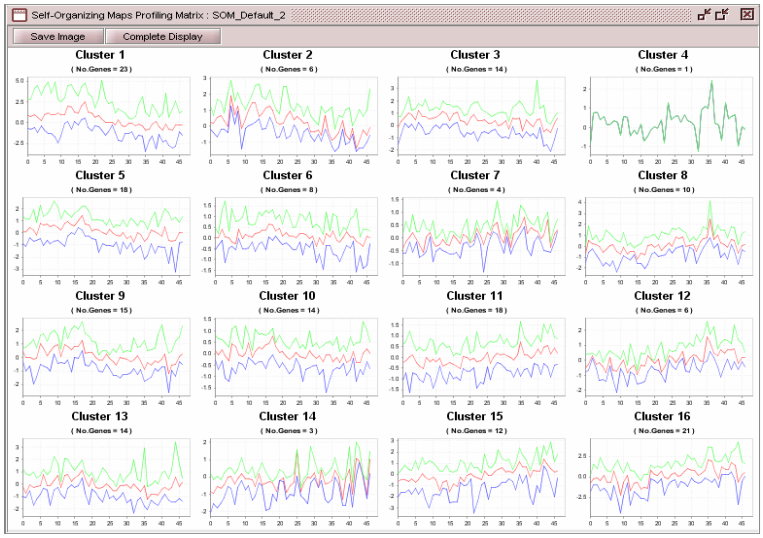
<Figure 4-19> U-Matrix (Distance View)

- Show Cluster Information: Can verify the information (Cluster order, Number of genes including in Cluster) of each Cluster.
- Show Similarity: Can verify the similarity between Clusters.

<Figure 4-20> U-Matrix (Show cluster Information, Show Similarity)

- Save Image: Possible to save U-matrix in picture file.

■ Profiling Matrix



<Figure 4-21> Self Organizing Map Profiling Matrix

① Save Image: It is possible to save Profiling Matrix in picture file.

② Complete Display: It shows entire gene profiling in the graph of each Cluster. Initial graph set up will only show Maximum, Median, Minimum.

■ Entire Cluster Result Window

<Figure 4-22> Entire Cluster Result Window

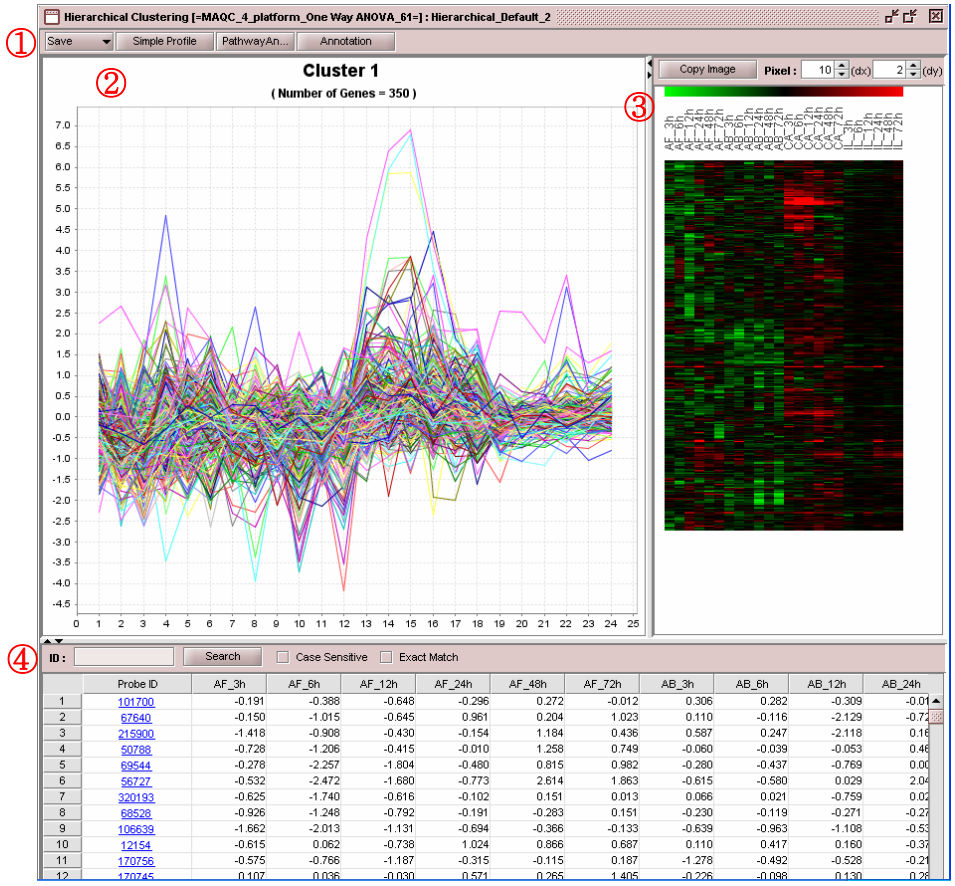
① Profiling Graph Range: It shows the profile (only Maximum, Median, Minimum) of each Cluster.

② Heatmap Range: It shows the Heat Map of used data of Clustering.

- Heatmap Pop Up Menu
  - Heatmap+ Annotation: On the right side of the Heatmap, user can see the Annotation Information in one screen. But Illumina Probe ID excluded.
  - Color: Scale around row Average: It is possible to control color of Up/Down based on the average of each gene.
  - Color: Yellow/Blue(up/down): It is possible to change the color of Heatmap in to Yellow and Blue.
  - Color: Brightness Scale: It is possible to control the brightness of Heatmap.
  - Copy image to Clipboard: It is possible to copy Heatmap image in to clipboard.
  - Save Image: It is possible to save Heatmat in to picture file.

③ Data Range: It shows the Cluster Order and Signal Intensity value among each genes in form of table. Click hyperlink of each ID to be connected to the URL database of the corresponding gene to verify detailed information of the corresponding gene.

■ Result Window of each Cluster



	Probe ID	AF_3h	AF_6h	AF_12h	AF_24h	AF_48h	AF_72h	AB_3h	AB_6h	AB_12h	AB_24h
1	101700	-0.191	-0.388	-0.648	-0.296	0.272	-0.012	0.306	0.282	-0.309	-0.01
2	67640	-0.150	-1.015	-0.645	0.961	0.204	1.023	0.110	-0.116	-2.129	-0.72
3	215900	-1.418	-0.908	-0.430	-0.154	1.184	0.436	0.587	0.247	-2.118	0.16
4	50788	-0.728	-1.206	-0.415	-0.010	1.258	0.749	-0.060	-0.039	-0.053	0.46
5	69544	-0.278	-2.257	-1.804	-0.480	0.815	0.982	-0.280	-0.437	-0.769	0.00
6	56727	-0.532	-2.472	-1.680	-0.773	2.614	1.863	-0.615	-0.580	0.029	2.04
7	320193	-0.625	-1.740	-0.616	-0.102	0.151	0.013	0.066	0.021	-0.759	0.02
8	68528	-0.926	-1.248	-0.792	-0.191	-0.283	0.151	-0.230	-0.119	-0.271	-0.23
9	106639	-1.662	-2.013	-1.131	-0.694	-0.366	-0.133	-0.639	-0.963	-1.108	-0.53
10	12154	-0.615	0.062	-0.738	1.024	0.866	0.687	0.110	0.417	0.160	-0.32
11	170756	-0.575	-0.766	-1.187	-0.315	-0.115	0.187	-1.278	-0.492	-0.528	-0.21
12	170745	0.107	0.036	-0.030	0.571	0.265	1.405	-0.226	-0.098	0.130	0.28

<Figure 4-23> Result Window of each Cluster

① Menu Bar on the Top

- Click [Save] button to save the information of corresponding Cluster.
  - [Matrix]: save gene data (ID and Signal intensity) among corresponding Cluster in to text file.
  - [Profile]: save Profiling graph of corresponding Cluster in to picture file.
  - [Heatmap]: save Heat Map image of corresponding Cluster in to picture file.
- Click [Full Profile] button to see all profiling of gene in Profiling graph of corresponding Cluster, it will be changed in [Simple Profile] to transform flexibly the graph of the Simple Profile and Full Profile.
- Click [Annotation] button to verify in form of table the Annotation Information of genes among corresponding Cluster.

- Pathway Analysis: Export genes of corresponding Cluster to the Pathway Analysis module.
- Annotation: Show the Annotation information of the genes of corresponding Custer.

② Profiling Graph Range: It shows the graph of corresponding Cluster Profiling. Genes of corresponding Cluster that the significant value is maximum, is marked in green, median in red, and minimum in blue.

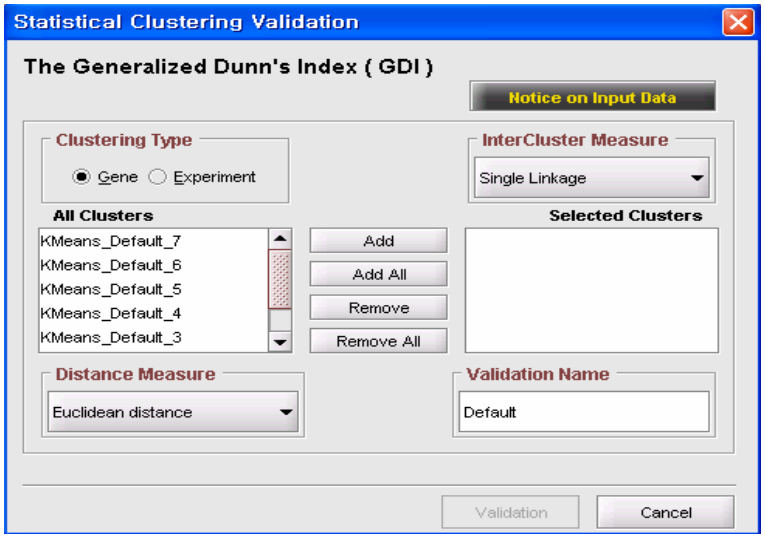
③ Heatmap Range: It shows the Heatmap of gene among corresponding Cluster.

④ Data Range: Data Range.

4.4. Validation

4.4.1. GDI

For GDI (The Generalized Dunn's Index) [4-5], click [Validation] → [GDI] on the menu bar, or click on the eleventh icon to see the set up window.



<Figure 4-24> The Generalized Dunn's Index (GDI) Set Up Window

① Clustering Type: Select standard format of comparing clustering result.
- Gene: Standardize the gene.
- Experiment: Standardize the Sample.

② InterCluster Measure: Select the method of calculating the linkage.
- Single Linkage
- Complete Linkage
- Average Linkage
- Centroid Linkage
- Average to Centroids
- Hausdorff
- All Linkage

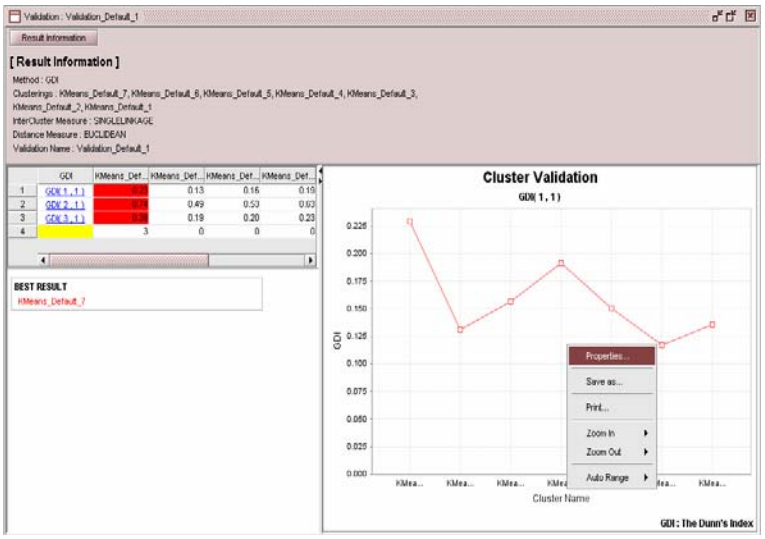
③ Click [Add] button on the list of left side of the window to select comparing Clustering Result.

④ Distance Measure: Select the method of calculation of the distance between Clusters (Ref.: 8.1).
- Euclidean Distance: Geometrical distance between two individuals

- Manhattan Distance: Distance between two individuals considering the gravity occupied in each variable.
- Pearson Correlation (centered): Measures the similarity grade of two individuals using the correlation coefficient after transforming each individual's average 0, and diversity 1.
- Pearson Correlation (uncentered): Measures the similarity grade of two individuals using actual signal value calculated correlation coefficient of two individuals.
- Absolute Pearson: Using absolute value of Pearson correlation coefficient.

⑤ GDI Input the name of the GDI result, and click [Validation] button to verify the result.

■ GDI 결과



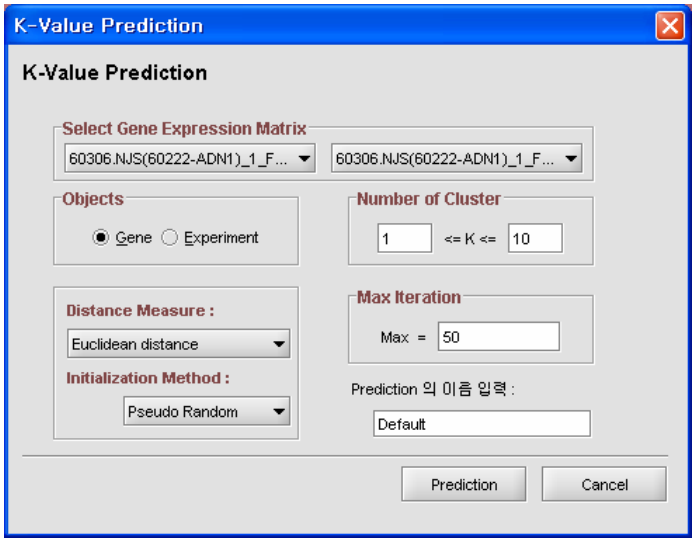
<Figure 4-25> GDI Result Window

① It is possible to verify the GDI detailed result from the left side table, and shows the name of the best rest on the bottom of the table. The result which has higher score than other Clustering result is marked in red cell on the table.

② It shows the GDI result on the right side range in graph format.

4.4.2. K-value Prediction

For K-value Prediction [4-6], select [Validation] → [K-value Prediction] on the menu bar, or click on the twelfth icon to see the set up window.



<Figure 4-26> K-value Prediction Set Up Window

① Select Gene Expression Matrix: Select the GEM to apply Clustering.

② Objects: Select the standard to apply Clustering.

- Gene: Standardize with gene.
- Experiment: Standardize with Sample.

③ Distance Measure: Select the method to calculate the distance (Ref.: 8.1).

- Euclidean Distance: Geometrical distance between two individuals.
- Manhattan Distance: The distance between two individuals considering the gravity occupying each variable.
- Pearson Correlation (centered): Measures the similarity grade of two individuals using the correlation coefficient after transforming each individual's average 0, and diversity 1.
- Pearson Correlation (uncentered): Measure the similarity grade of two individuals using actual signal value calculated correlation coefficient of two individuals.
- Absolute Pearson: Using absolute value of Pearson correlation coefficient

④ Initialization Method: Select initializing method.

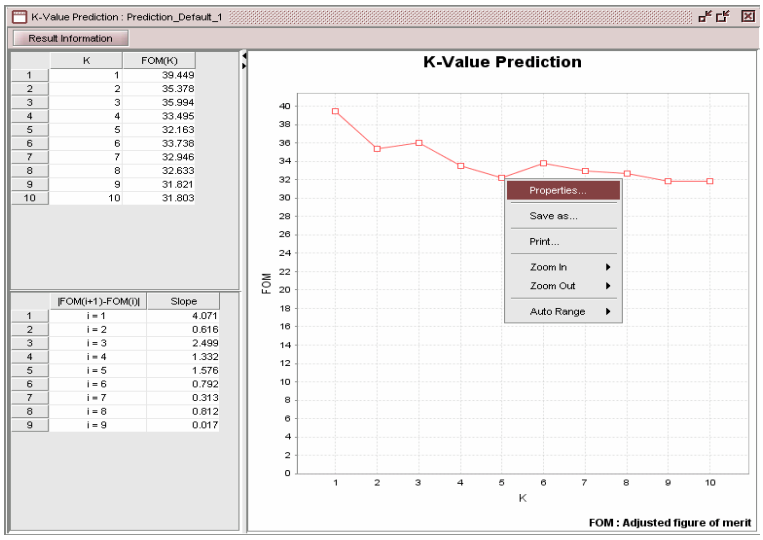
- Pseudo Random: The method generating similar random number in every repetition.
- Totally Random: The method generating optional random number in every repetition.

⑤ Number of Cluster: Input the range of predicted Cluster Number (K).

⑥ Max Iteration: 최대 Input maximum repetition frequency (Basic value: 50).

⑦ The name of the Prediction Result, and click [Prediction] button to verify the result.

■ Prediction Result



<Figure 4-27> K-Value Prediction Result Window

① It is possible to verify the result of FOM according the K value from the left side table.

② It shows the result in graph format on the right side range.

4.5. Reference

[4-1] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA, 95, 14863-14868.

[4-2] J.A. Hartigan & M.A. Wong (1979) A k-means clustering algorithm. Appl. Statist. 28:100-108.

[4-3] S. Tavazoie et al. (1999) Systematic determination of genetic network architecture. Nat. Genet., 22, 281-285.

[4-4] P. Tamayo et al. (1999) Interpreting patterns of gene expression with SOMs – methods and application to hematopoietic differentiation. Proc. Natl Acad. Sci. USA, 96, 2907-2912.

[4-5] F. Azuaje (2002) A cluster validity framework for genome expression data. Bioinformatics, 18, 319-320.

[4-6] K.Y. Yeung et al. (2001) Validating clustering for gene expression data. Bioinformatics, 17, 309-318.

Classification

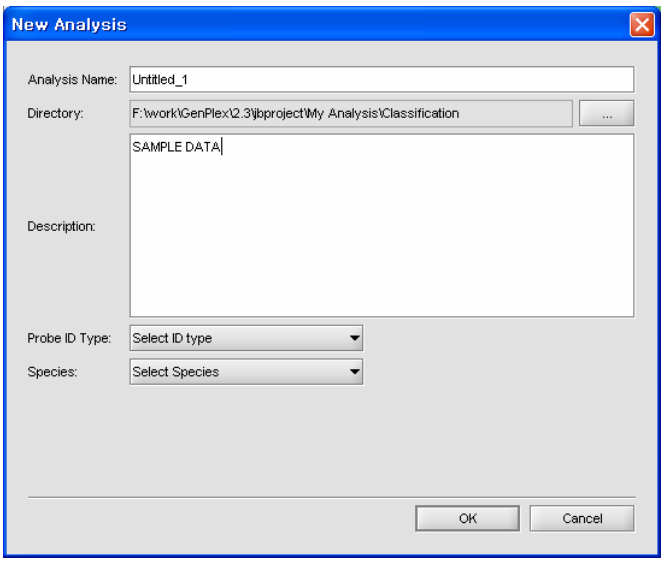
5. Classification

5.1. File

5.1.1. New Analysis

In case of exporting data from Preprocessing module, Analysis file is automatically created, so this procedure can be skipped (Go to 5.2 ▷).

On the menu bar, select [File] → [New Analysis], or click on the first icon, then the Analysis creating window appears.



<Figure 5-1> Analysis Creating Window

① Analysis Name: Input the name of the Analysis File.

② Directory: Click [⋯] button to select the position where Analysis File will be created.

③ Description: Input the additional information related to Analysis (can be skipped).

④ Probe ID Type: Select the ID Type of the input data.

- Commercial Product Probe ID
  - Affymetrix GeneChip Probe ID
  - Agilent Probe ID(One-dye)
  - Agilent Probe ID(Two-dye)
  - Applied Biosystems 1700 Probe ID

- CodeLink Probe ID
- Illumina Probe ID
- Operon Probe ID

- Public DataBase ID
  - IMAGE Clone ID
  - NCBI Clone ID
  - NCBI GenBank Accession
  - NCBI GeneID (LocusLink)
  - NCBI UniGene ID
- Others (In case ID unknown)

⑤ Species: Select the species of the input data.

⑥ Click [OK] button to have Analysis File created, then the data selecting window will appear (Go to 5.1.7 ▷).

5.1.2. Open Analysis

On the menu bar, select [File] → [Open Analysis], or click on the second icon to open saved Analysis File.

5.1.3. Recent Analysis

On the menu bar, select [File] → [Recent Analysis] to open recently analyzed Analysis File. This list can be deleted using [Clear History] menu.

5.1.4. Save Analysis

On the menu bar, select [File] → [Save Analysis], or click on the third icon to save the working Analysis File.

5.1.5. Save Analysis As

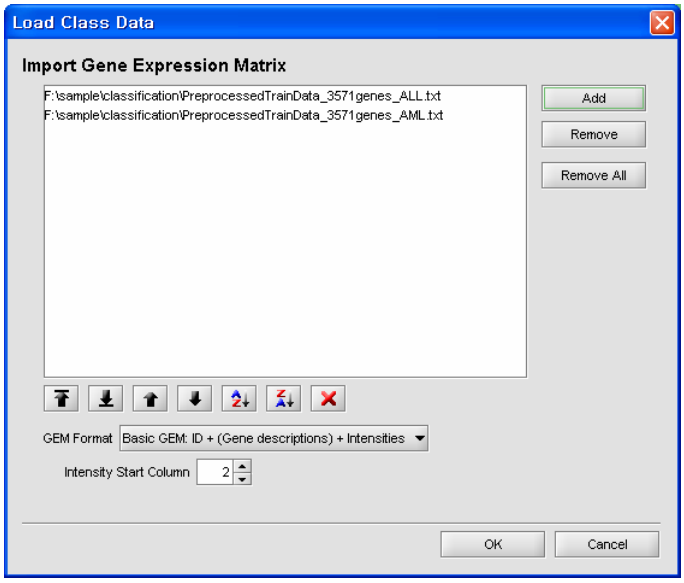
On the menu bar, select [File] → [Save Analysis As...], or click on the forth icon to save working Analysis File in a different file name.

5.1.6. Close Analysis

On the menu bar, select [File] → [Close Analysis] to close working Analysis File.

5.1.7. Load Training Data File(s)

On the menu bar, select [File] → [Load Training Data File(s)], or click on the fifth icon , then the data selecting window will appear.



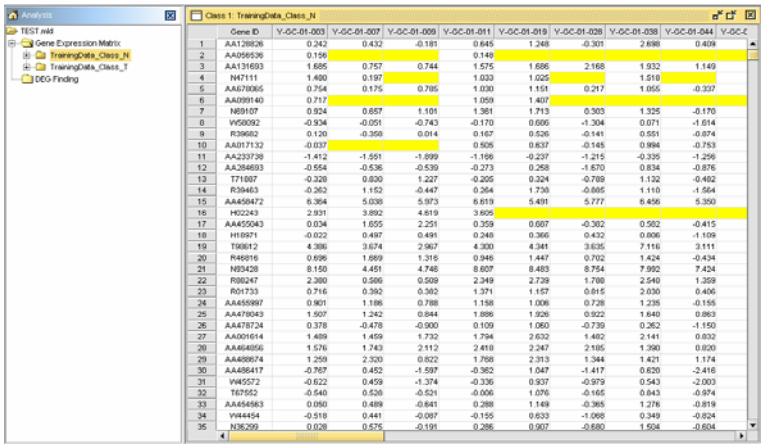
<Figure 5-2> Class Data Selecting Window

① Click [Add] button to add inputting file, it will be added on the list on left side of the window, and use [Remove] and [Remove All] button to delete the item.

② Condition Column Start Position: Fix the position where intensity column in input data.

③ Click [Check] button to input data.

■ Input Data Result



<Figure 5-3> Input Data Verifying Window

① Double click the input data name on the browse window to verify input data. Missing Values are marked in yellow.

5.1.8. Load Test Data File(s)

On the menu bar, select [File] → [Load Test Data File(s)], or click on the sixth icon , then data selecting window will appear (Ref.: 5.1.7).

5.1.9. Transpose Data

On the menu bar, select [File] → [Transpose Data] to transform the Row and the Column of the input data.

5.1.10. Analysis Properties

On the menu bar, select [File] → [Analysis Properties] to adjust the information of working Analysis File. 5.1.11. Exit On the menu bar, select [File] → [Exit] to complete the program.

5.1.11. Exit

On the menu bar select [File] → [Exit], then you can finish the program.

5.2. Preprocessing

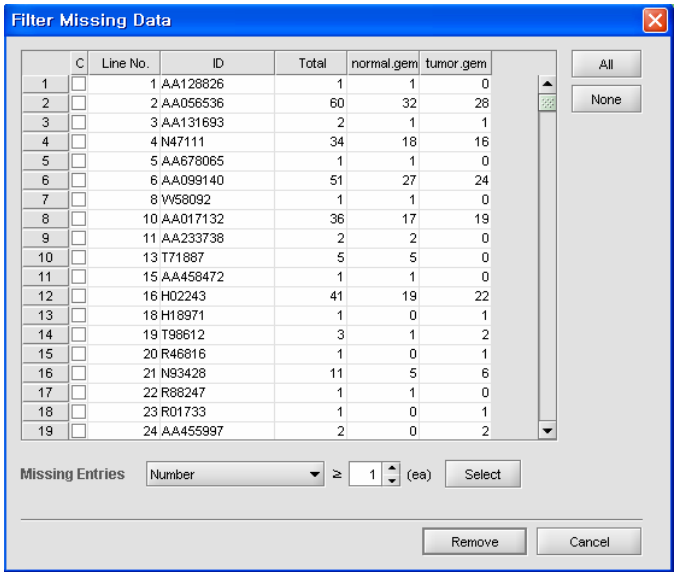
In the Preprocessing menu, it is possible to verify whether input data format matches or delete Missing Data. Then, to continue the analysis, input data format has to match and no Missing Data.

5.2.1. Check & Match Data

On the menu bar, select [Preprocessing] → [Check & Match Data] to verify whether input data gene figure and ID is identical. In case each input data gene figure or ID does not match, it is possible to match based on gene ID.

5.2.2. Filter Missing Data

On the menu bar, select [Preprocessing] → [Filter Missing Data], then the window as seen below will appear, and it is possible to delete missing entry from input data.



<Figure 5-4> Filter Missing Data Input Window

① Missing Entries: Select the standard and click [Select] button to select proper genes to be deleted.

- Number: Select genes to be deleted based on the Total Column of the Missing Value.
- Percentage: Select genes to be deleted based on the Ratio of Total Number of sample and Total Column of Missing Value.

② Click [Remove] button to delete all selected missing entry from Training Data and Test Data.

5.2.3. Impute Data

On the menu bar, select [Preprocessing] → [Impute Data] for the user to complete the missing value according to regular rule.

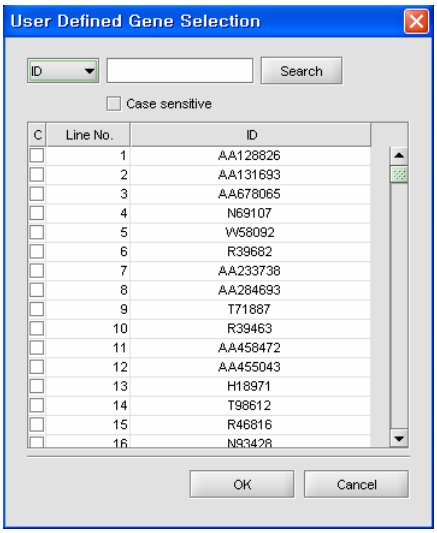
5.3. Gene Selection

Select Marker Gene from input data using statistical method. Gene Selection is possible if only there are more than two Training Data.

5.3.1. Select Algorithm

On the menu bar, select [Gene Selection] → [Select Algorithm...], or click on the seventh icon to select wanted Algorithm.

- Gene Selection Algorithm
  - Null: Select entire gene (Basic value).
  - Two-sample t-test: Only possible to select when there are two Training data.
  - BSS/WSS: cluster Use the error between clusters and ratio of error among cluster [5-1].
  - Kruskal-Wallis H-Test: Method to compare more than three cluster distribution.
  - Regularized t-test
  - User defined: User can directly define significant genes.



<Figure 5-5> User Defined Gene Selection Communication Window

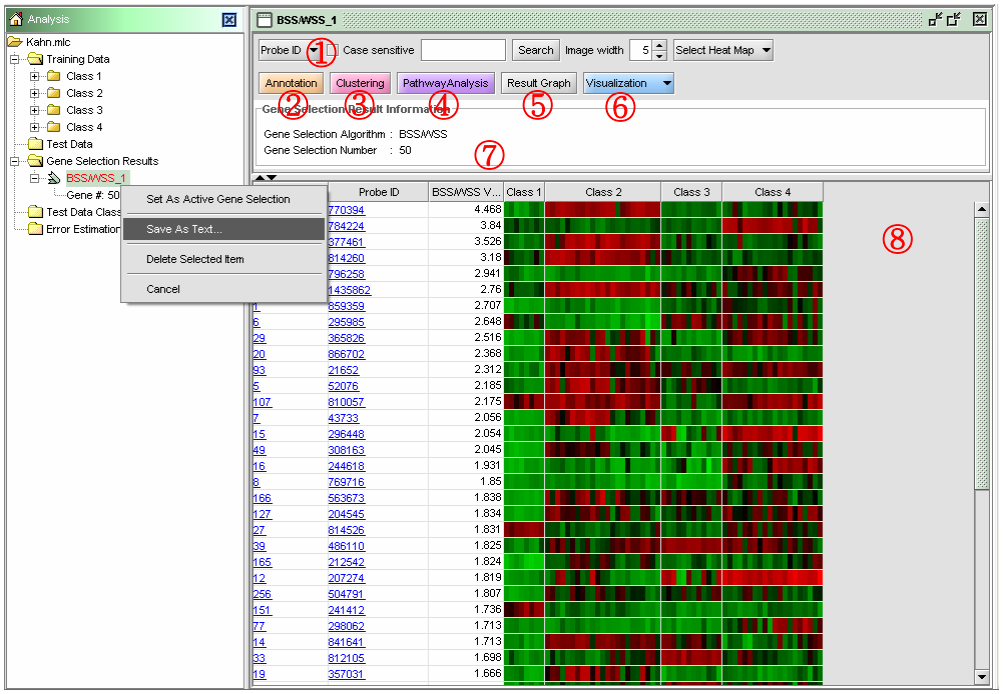
5.3.2. Set Parameter(s)

It is possible to set up the parameter (number of genes or p-value) of Gene Selection Algorithm that user have selected, if not set up, basic value will be used. If Null or

User defined is selected as Gene Selection Algorithm, then this menu will inactivate because there is no parameter. On the menu bar, select [Gene Selection] → [Set Parameter(s)...], or just click on the eighth icon .

5.3.3. Run

On the menu bar, select [Gene Selection] → [Run], or click on the ninth icon to see the Gene Selected Result which the user have selected and set up an Algorithm and Parameter. The last operation result will be fixed to basic Gene Selection Result and will be used to Test data distinction.



<Figure 5-6> Gene Selection Result Window

① Browse Engine and Heat Map Set Up

- Search: Search with ID or Line No. Select search type, and input the ID or Line No. on the text window and click [Search] button, then the corresponding gene will be shown reversed on the result table. Click [Case Sensitive] to search classified by the Capital Letter and the Small Letter of alphabet.
- Image width: Change the width of the Heat Map shown on the Result window. Input the width to be changed and just click [Enter].

- Select Heat Map: The color of Heat Map will be changed.
  - Red/Green
  - Blue/Yellow

② Annotation: Can verify the Annotation Information on the selected genes (Ref.: 3.3.1.3)

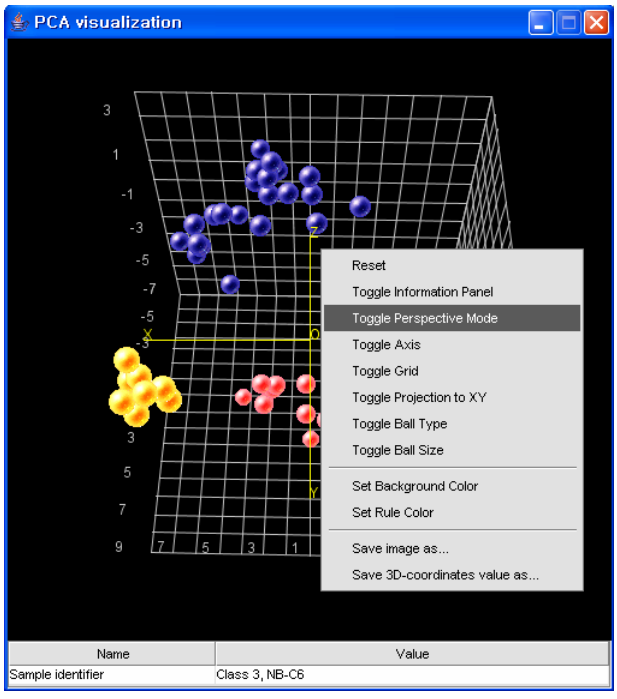
③ Clustering: Export selected gene data only to the Clustering module.

④ Pathway Analysis: Export selected gene data only to the Pathway Analysis module.

⑤ Result Graph: Gene Selection Result will be shown in the graph. X axis is the Rank, Y axis is the Result Value of the calculation. Drag the mouse point following the graph, then the user can verify the rank and the result value of the calculation.

⑥ Visualization: Can verify visually in three-dimension of Training Data distinctive or not of the Gene Selection Result.

- 3-Gene based: It shows in 3D using only high ranked 3 genes among Gene Selection Result.
- PCA: It shows PCA Result in 3D, using all Gene Selection Result.
- Each ball shows Training Data or Test Data of each sample and when user selects this ball, the Sample Information will be shown on the table below.

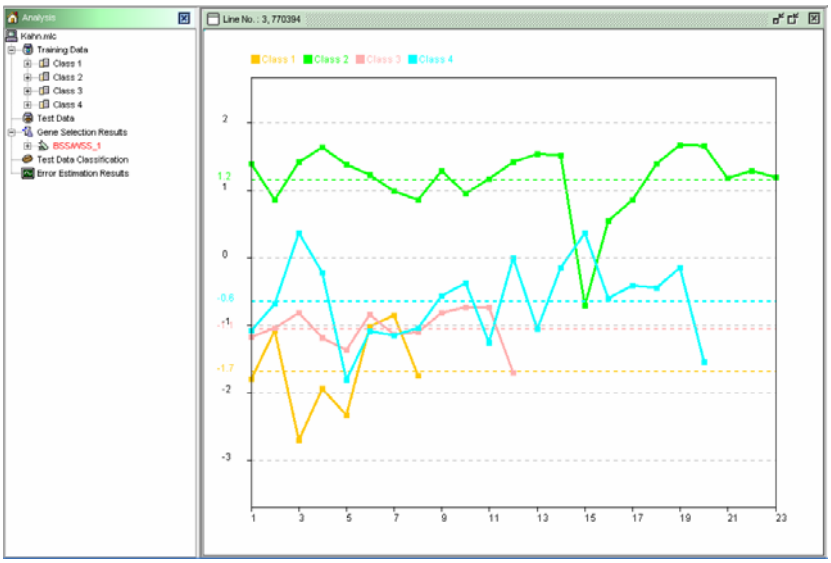


<Figure 5-7> Visualization Result Window

⑦ It shows set up of Gene Selection Algorithm and Parameter that user have selected.

⑧ It shows the result of Gene Selection.

- Line No: It shows the rank of input data genes. Click hyperlink to see the gene profile graph. Gene profile graph shows the average of study data in dotted line, and shows corresponding gene significant pattern in bended line graph.



<Figure 5-8> Gene Profile Result Window

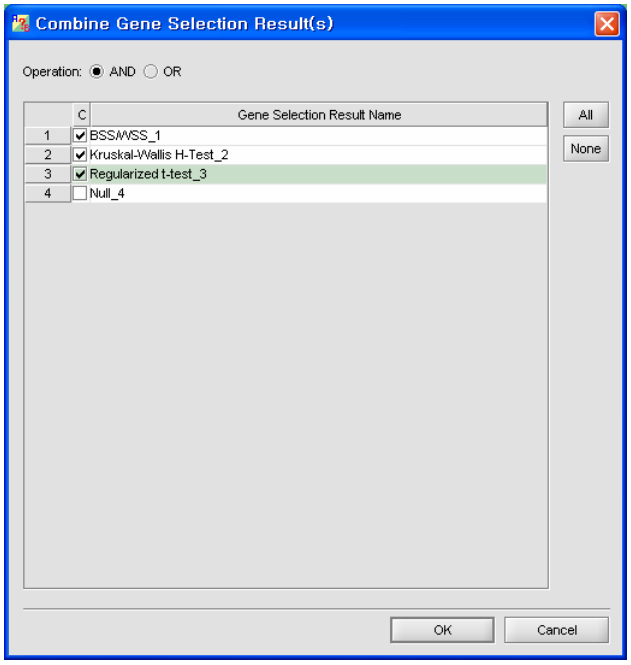
- ID: It shows the ID of each gene. Click hyperlink to verify the detailed information of corresponding gene in connecting related database URL. To have exact information, exact ID Type should be selected when Analysis is created. When ID Type is selected as Other or All, then it will not be connected.
- Score: It is the calculation result value of each gene. This value shows the variation progress in the result graph, so it can be used to predict visually the gene which shows the variation of significant value.
- It shows the Heat Map of the extracted significant genes. User can easily verify with eyes the significant information of each gene.

5.3.4. Combine Results

On the menu bar, select [Gene Selection] → [Combine Results], or click on the tenth

icon to associate several Gene Selection Result using AND, OR operation.

If there is already a Marker gene, or if it is needed to use certain genes which, the biological information are known for useful distinction, use [User Defined] to select directly corresponding genes and possible to use in associating with existing Gene Selection result.



<Figure 5-9> Combined Gene Selection Communication Window

5.3.5. Set As Active Gene Selection

On the menu bar, select [Gene Selection] → [Set As Active Gene Selection], then it will be activated to Gene Selection Result used to Test data distinction.

5.3.6. Export to Clustering

On the menu bar, select [Gene Selection] → [Export to Clustering Module], then it is possible to export several Gene Selection results in to Clustering module.

5.3.7. Export to Pathway Analysis Module

On the menu bar, select [Gene Selection] → [Export to Pathway Analysis Module], then it is possible to export Gene Selection result in to Pathway Analysis module.

5.3.8. Save Result(s)

On the menu bar, select [Gene Selection] → [Save Result(s)...] to save selected Gene Selection result in to text file.

5.4. Classification

It is possible to distinct the Test Data using Gene Selection result. Classification is possible if only there are more than two Training Data and more than one Test Data.

5.4.1. Select Distance

In Classification, distance calculation between vectors is used, and the user can select the method of distance calculation at this point. On the menu bar, just select [Classification] → [Select Distance...] (Basic value: Euclidean Distance - Ordinary).

- Classification Distance (Ref.: 8.1)
  - Euclidean Distance
    - Ordinary: The method using the geometrical distance between two vectors
    - SD-weight: Use the calculated distance with weight with standard deviation between two vectors
  - Manhattan Distance: Calculate considering the ratio of each variation occupying.
  - Minkowski Distance
    - 3~9
  - Pearson Correlation Coefficient: The method using the Correlation Coefficient of two vectors.

5.4.2. Select Algorithm

On the menu bar, select [Classification] → [Select Algorithm...], or just click on the eleventh icon (Basic value: Weighted K-Nearest Neighbor).

- Classification Algorithm
  - Weighted K-Nearest Neighbor: Decide the class of the given individual considering the class that this K unit of individual belongs, after calculating nearest K unit of individual with given individual.
  - Prototype Matching with indeterminacy parameters.
  - Multi-FLDA: The method to assign to class forming the linear distinction.

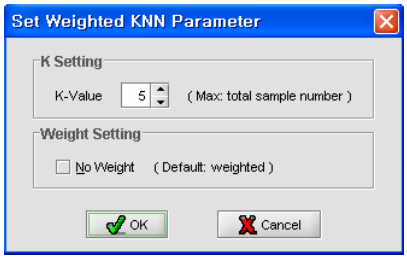
5.4.3. Set Parameter(s)

It is possible to set up the parameter of Classification Algorithm that the user has selected, but if not set up basic value will be used. In case of Classification Algorithm and selected Multi-FLDA is selected, corresponding menu will inactivate because there is no parameter.

On the menu bar, select [Classification] → [Set Parameter(s)…], or just click on the twelfth icon .

■ Weighted K-Nearest Neighbor (KNN)

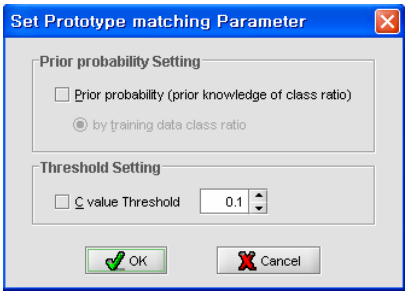
Select whether using K value and weight or not (Basic value: K=5, weighted).



<Figure 5-10> Weighted KNN Parameter Input Window

■ Prototype Matching with indeterminacy parameters

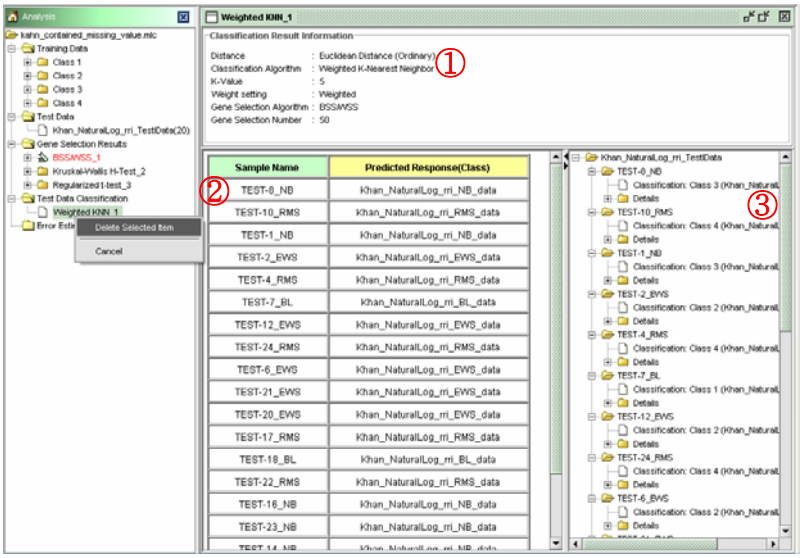
If the calculation result is under designated C value, it is determined as indeterminate (Basic value: C=0.1).



<Figure 5-11> Prototype Matching Parameter Input Window

5.4.4. Classify Test Data

On the menu bar, select [Classification] → [Classify Test Data], or click on the thirteenth icon to verify the Classification result.



<Figure 5-12> Classification Result Window

① It shows Classification Algorithm and Parameter set up information.

② It is easy to verify the distinguished result on each sample of Test data in table form.

③ It provides detailed information of distinguished result in tree format.

5.5. Error Estimation

It can measure the Error Estimation using Gene Selection set up and Classification set up that the user have selected. Error Estimation is possible only if there are more than two Training Data.

5.5.1. Select Algorithm

On the menu bar, select [Error Estimation] → [Select Algorithm...], or just click on the fourteenth icon (Basic value: LOOCV).

- Error Estimation Algorithm[5-3]
  - LOOCV: This is the method when K=n in K-fold method.
  - K-Fold: Divide the data into K unit of fold. Use K-1 unit as training set and another one as test set to sort out Error Estimation of K times and then calculate misclassification rate.
  - Bootstrap: bootstrap Calculate misclassification rate through bootstrap sampling.

5.5.2. Set Parameter(s)

It is possible to set up the parameter of Error Estimation Algorithm that user have selected. If it is not set up, Basic Value will be used.

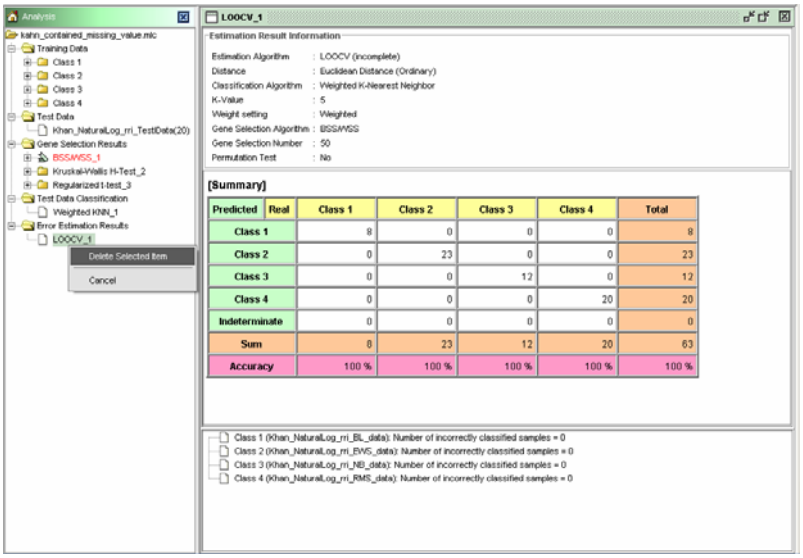
On the menu bar, select [Error Estimation] → [Set Parameter(s)…], or just click on the fifteenth icon .

- LOOCV (Basic Value: Incomplete)
- K-Fold (Basic Value: Incomplete, Fold Number=10, Iteration Number=100)
- Bootstrap (Basic Value: B=50)

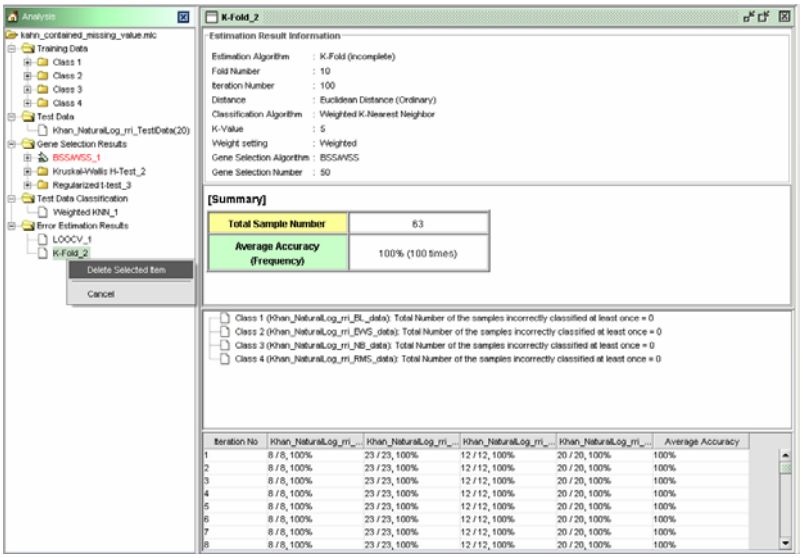
5.5.3. Run

On the menu bar, select [Error Estimation] → [Run], or click on the sixteenth icon to verify the Error Estimation Result.

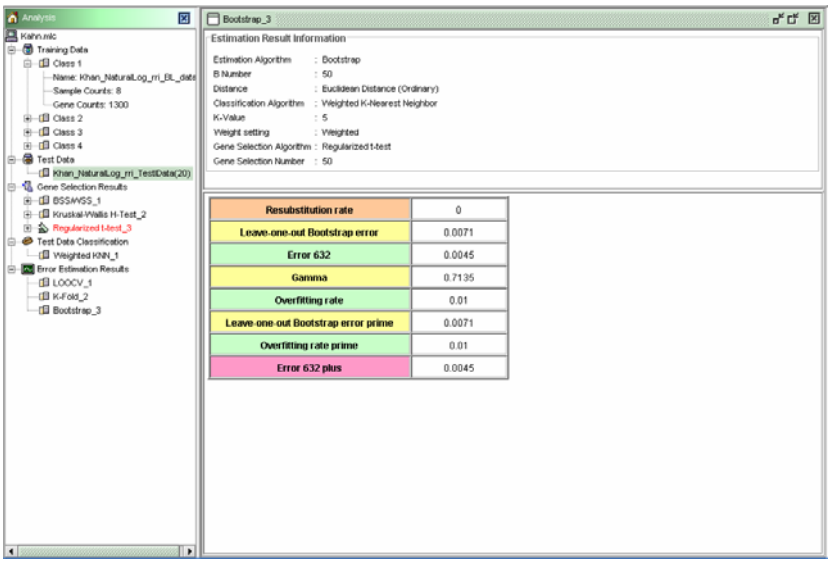
It is possible to verify Error Estimation Algorithm and Parameter set up information and detailed information of the result through result window.


Analysis
khan_contained_missing_value.mlc
Training Data
Class 1
Class 2
Class 3
Class 4
Test Data
Khan_NaturalLog_rri_TestData(20)
Gene Selection Results
BSSAWSS_1
Kruskal-Wallis H-Test_2
Regularized t-test_3
Test Data Classification
Weighted KNN_1
Error Estimation Results
LOOCV_1
Delete Selected Item
Cancel
LOOCV_1
Estimation Result Information
Estimation Algorithm : LOOCV (incomplete)
Distance : Euclidean Distance (Ordinary)
Classification Algorithm : Weighted K-Nearest Neighbor
K-Value : 5
Weight setting : Weighted
Gene Selection Algorithm : BSSAWSS
Gene Selection Number : 50
Permutation Test : No
[Summary]
Predicted | Real | Class 1 | Class 2 | Class 3 | Class 4 | Total
Class 1 | 8 | 0 | 0 | 0 | 8
Class 2 | 0 | 23 | 0 | 0 | 23
Class 3 | 0 | 0 | 12 | 0 | 12
Class 4 | 0 | 0 | 0 | 20 | 20
Indeterminate | 0 | 0 | 0 | 0 | 0
Sum | 8 | 23 | 12 | 20 | 63
Accuracy | 100 % | 100 % | 100 % | 100 % | 100 %
Class 1 (Khan_NaturalLog_rri_BL_data): Number of incorrectly classified samples = 0
Class 2 (Khan_NaturalLog_rri_EWS_data): Number of incorrectly classified samples = 0
Class 3 (Khan_NaturalLog_rri_NB_data): Number of incorrectly classified samples = 0
Class 4 (Khan_NaturalLog_rri_RMS_data): Number of incorrectly classified samples = 0

<Figure 5-13> Error Estimation Result Window (LOOVC)


Analysis
khan_contained_missing_value.mlc
Training Data
Class 1
Class 2
Class 3
Class 4
Test Data
Khan_NaturalLog_rri_TestData(20)
Gene Selection Results
BSSAWSS_1
Kruskal-Wallis H-Test_2
Regularized t-test_3
Test Data Classification
Weighted KNN_1
Error Estimation Results
LOOCV_1
K-Fold_2
Delete Selected Item
Cancel
K-Fold_2
Estimation Result Information
Estimation Algorithm : K-Fold (incomplete)
Fold Number : 10
Iteration Number : 100
Distance : Euclidean Distance (Ordinary)
Classification Algorithm : Weighted K-Nearest Neighbor
K-Value : 5
Weight setting : Weighted
Gene Selection Algorithm : BSSAWSS
Gene Selection Number : 50
[Summary]
Total Sample Number | 63
Average Accuracy (Frequency) | 100% (100 times)
Class 1 (Khan_NaturalLog_rri_BL_data): Total Number of the samples incorrectly classified at least once = 0
Class 2 (Khan_NaturalLog_rri_EWS_data): Total Number of the samples incorrectly classified at least once = 0
Class 3 (Khan_NaturalLog_rri_NB_data): Total Number of the samples incorrectly classified at least once = 0
Class 4 (Khan_NaturalLog_rri_RMS_data): Total Number of the samples incorrectly classified at least once = 0
Iteration No | Khan_NaturalLog_rri_... | Khan_NaturalLog_rri_... | Khan_NaturalLog_rri_... | Khan_NaturalLog_rri_... | Average Accuracy
1 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%
2 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%
3 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%
4 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%
5 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%
6 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%
7 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%
8 | 8 / 8, 100% | 23 / 23, 100% | 12 / 12, 100% | 20 / 20, 100% | 100%

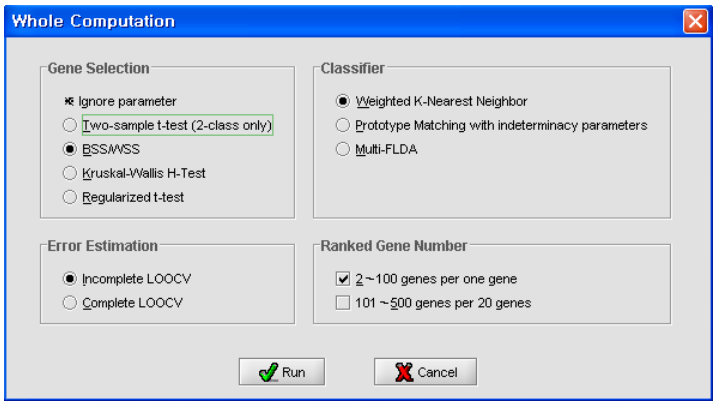
<Figure 5-14> Error Estimation Result Window (K-Fold)



<Figure 5-15> Error Estimation Result Window (Bootstrap)

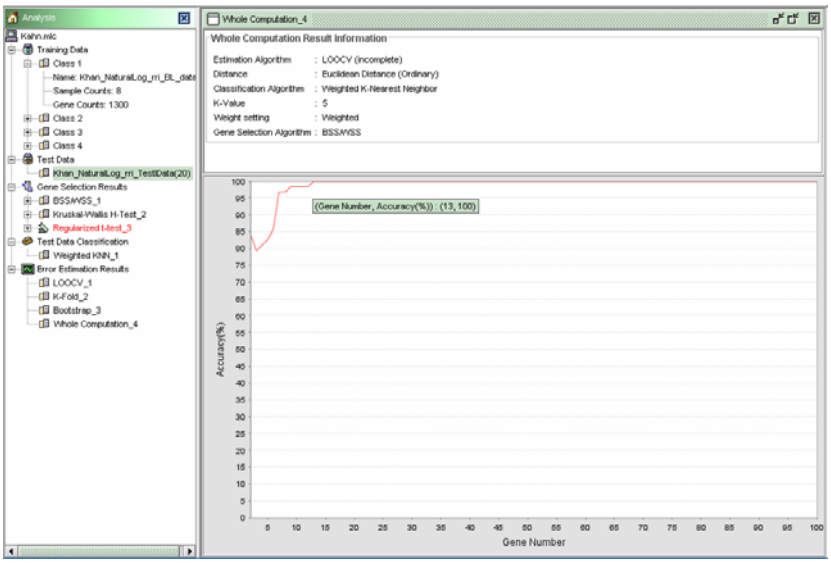
5.5.4. Whole Computation

On the menu bar, select [Error Estimation] → [Whole Computation], or click on the seventeenth icon , then the Whole Computation Set Up window appears.



<Figure 5-16> Whole Computation Set Up Window

Select each Algorithm of Gene Selection, Classification, Error Estimation, and click [Run] button to verify once for all the Error Estimation following number of genes. In the Whole Computation Result Window, drag the mouse point through the graph, then it is possible to verify number of each gene and Error Estimation, and also can save the result graph in to picture file.



<Figure 5-17> Whole Computation Result Window

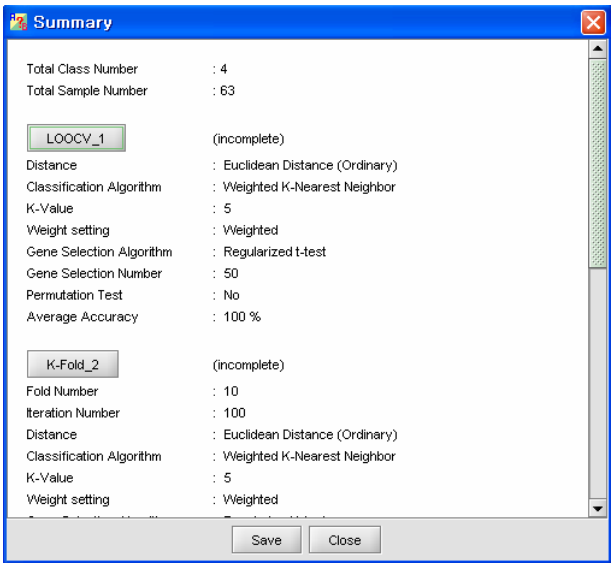
5.6. View

5.6.1. Show Sample 3D View

On the menu bar, select [View] → [Show Sample 3D View] to verify visually in 3 Dimension, whether the Training Data is distinctive or not of 3 genes that the user have designated.

5.6.2. Show Summary View

On the menu bar, select [View] → [Show Summary View] to verify the summarized information of Error Estimation and to save it in to text file.



<Figure 5-18> Error Estimation Result Summarizing Window

5.7. Reference

[5-1] S. Dudoit et al. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Stat. Association, 97, 77-87.

[5-2] R. Tibshirani et al. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl Acad. Sci. USA. 99, 6567-6572.

[5-3] C. Ambroise and G.J. McLachlan (2002) Selection bias in gene extraction on the basis of microarray gene expression data. Proc. Natl Acad. Sci. USA. 99, 6562-6566.

[5-4] R.L. Somorjai, B. Dolenko, R. Baumgartner (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics, 19(12):1484-1491.

Pathway Analysis

6. Pathway Analysis

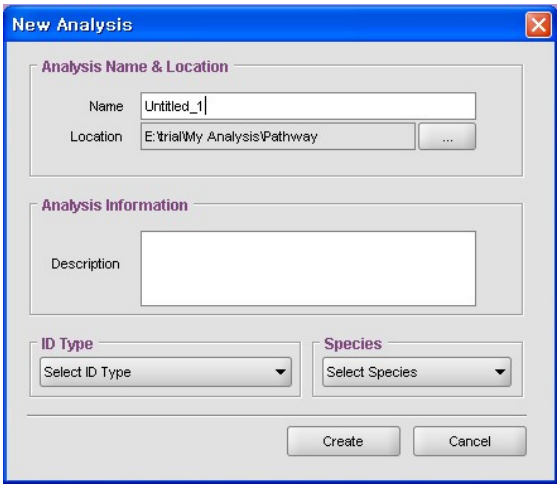
Pathway Analysis is the method to survey the Pathway Information of input data (the list of genes with significant value). Pathway Analysis Result provides easy biological interpretation applying various visual functions and editing function.

6.1. File

6.1.1. New Analysis

In case of the data is exported from other module, the Analysis File will be created automatically, so this procedure does not correspond (Go to 6.2 ▷).

On the menu bar, select [File] → [New Analysis], or click on the first icon , then Analysis creating window appears.



< Figure 6-1> Analysis Creating Window

① Name: Input the name of created Analysis File.

② Location: Click […] button to select the position where Analysis File will be created.

③ Description: Input added information in to the Analysis File (it can be skipped).

④ ID Type: Select the ID Type of the inputting data.

- Commercial Product Probe ID
  - Affymetrix GeneChip Probe ID
  - Agilent Probe ID(One-dye)
  - Agilent Probe ID(Two-dye)

- Applied Biosystems 1700 Probe ID
- CodeLink Probe ID
- Illumina Probe ID
- Operon Probe ID

● Public DataBase ID
- IMAGE Clone ID
- NCBI Clone ID
- NCBI GenBank Accession
- NCBI GeneID (LocusLink)
- NCBI UniGene ID

● Others (If ID not known)

⑤ Species: Select the species of input data.

⑥ Click [Create] button, then the Analysis File will be created and the data selecting window appears (Go to 6.1.7 ▷).

6.1.2. Open Analysis

On the menu bar, select [File] → [Open Analysis], or click on the second icon to open the saved Analysis File.

6.1.3. Recent Analysis

On the menu bar, select [File] → [Recent Analysis] to open recently analyzed Analysis File. This list can be deleted using [Clear History] menu.

6.1.4. Save Analysis

On the menu bar, select [File] → [Save Analysis] or click on the third icon to save working Analysis File.

6.1.5. Save Analysis As

On the menu bar, select [File] → [Save Analysis As...] or click on the fourth icon to save working Analysis File in different name.

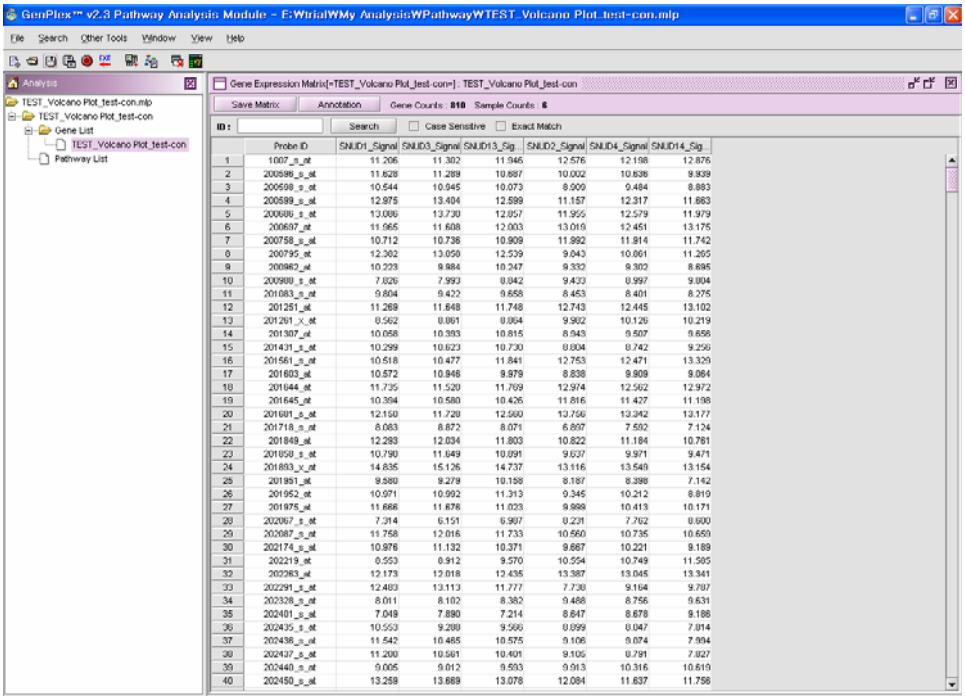
6.1.6. Close Analysis

On the menu bar, select [File] → [Close Analysis] to close working Analysis File.

6.1.7. Import Data

On the menu bar, select [File] → [Import Data], or click on the fifth icon, then the data selecting window will appear.

■ Data Input Result



< Figure 6-2> Input Data Verifying Window

① Double click the name of the input data to verify the input data on the searching window.

② Click the button on the upper side of the Data Verifying Window to save it in to text file, or can verify Annotation Information of corresponding genes.

6.1.8. Analysis Properties

On the menu bar, select [File] → [Properties Analysis] to adjust the information on

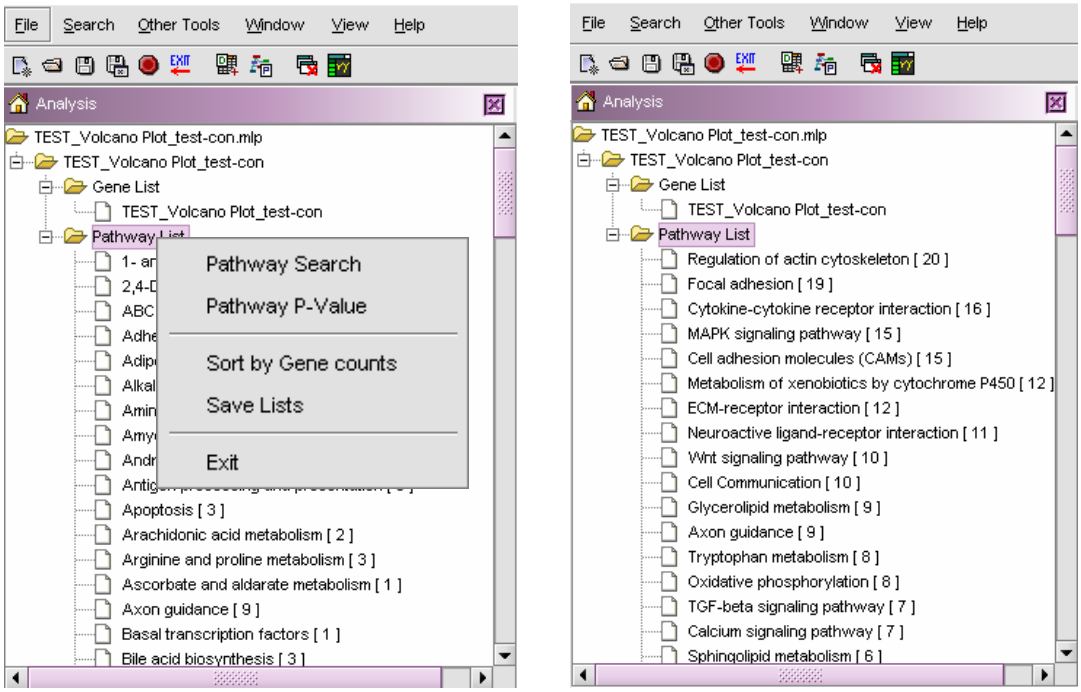
working Analysis File.

6.1.9. Exit

On the menu bar, select [File] → [Exit] to close the program.

6.2. Pathway List

① In the Pathway List folder, right click on the mouse [Pathway Search], [Pathway P-Value] menu, and then it is possible to verify KEGG Pathway corresponding input data and P-Value of the corresponding Pathway.


File
Search
Other Tools
Window
View
Help
Analysis
TEST_Volcano Plot_test-con.mlp
TEST_Volcano Plot_test-con
Gene List
TEST_Volcano Plot_test-con
Pathway List
Pathway Search
Pathway P-Value
Sort by Gene counts
Save Lists
Exit
Apoptosis [3]
Arachidonic acid metabolism [2]
Arginine and proline metabolism [3]
Ascorbate and aldarate metabolism [1]
Axon guidance [9]
Basal transcription factors [1]
Bile acid biosynthesis [3]
Regulation of actin cytoskeleton [20]
Focal adhesion [19]
Cytokine-cytokine receptor interaction [16]
MAPK signaling pathway [15]
Cell adhesion molecules (CAMs) [15]
Metabolism of xenobiotics by cytochrome P450 [12]
ECM-receptor interaction [12]
Neuroactive ligand-receptor interaction [11]
Wnt signaling pathway [10]
Cell Communication [10]
Glycerolipid metabolism [9]
Axon guidance [9]
Tryptophan metabolism [8]
Oxidative phosphorylation [8]
TGF-beta signaling pathway [7]
Calcium signaling pathway [7]
Sphingolipid metabolism [6]

<Figure 6-3> KEGG Pathway Search Result

② If you select [Sort by Gene counts] menu, it will show the Pathway list in bigger number order of the number of related genes in an array.

③ If you select [Save List] menu, it will save the Pathway list in text file.

6.2.1. Pathway (Image)

① Double click the name of Pathway on the left side tree to see the KEGG Pathway image.

② Genes related with corresponding Pathway is marked in red box.

③ Right click on the mouse to see the popup menu.

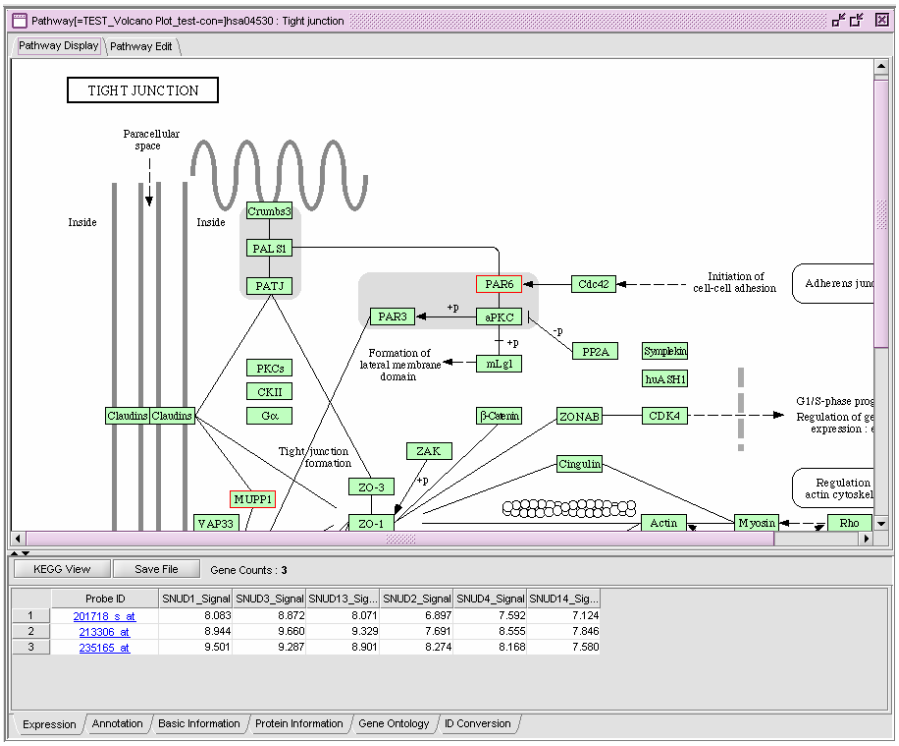
- Click [Hide Gene] to hide the marked related genes.
- Click [Show Heat Map] to see each Heat Map (expressed information) to the related genes.

- Click [Heatmap color: up/down] to fix the color of the gene expressed value.
- Click [Save Image] to save as picture file.

④ Below, it provides related gene Signal and Annotation Information.

6.2.2. Pathway (XML)

① Simple editing is possible in [Pathway (XML)] tab.



<Figure 6-4> Pathway Map Figure

Algorithms

7. DEF Finding Algorithm

This is the method called DEG (Differentially Expressed Gene) Finding, that is to find out genes expressed differently in statistics between analysis group (e.g.: compare between control group and treatment group).

7.1. Fold Change

This method was mainly used in early days of DNA chip analysis, because of its strong points which is, simplicity in applying and easy interpretation of result. These are generally used until recent days.

Calculate the significant figure between control sample (reference sample) and treatment sample of each gene, and then it is to see how much the treatment sample expressed relatively compared to the control sample. Generally, fold change is known as the ratio value itself, but sometimes value transformed in Log_2 format is also called fold change (for convenience, we will understand fold change as transformed in Log value, hereafter).

For Fold Change, we have to set up the threshold of ratio value to obtain DEG sampling, generally 2 fold is the standard, it can be lowered to 1.5 fold or raised up to more than 4 fold according to the data.

But, it can be a problem applying this kind of batch processing. For example, when 2 fold is applied, there are relatively more genes satisfying corresponding condition in low expressed region. But on the other side it is hard to satisfy 2 fold condition in high expressed region.

Also fold change does not consider statistical significance of variance among the group of gene expressed figure when comparing between groups. For example, if 3 control class samples and 3 experiment class samples are given, we take an average calculating total 9 case of fold change of each gene to have DEG in fold change method. But it is hard to say that this average value represents all 9 cases without mentioning how much we trust on the statistics.

Because of it can be distorted, even if there are one or two outliers among these 9 cases. For these reasons, we can figure out that the fold change is more an experimental method than the statistical method

7.2. Two-sample (unpaired) t-test

This method is broadly used together with Fold Change in obtaining DEG, but it is contrary to the Fold Change because it gives statistical significance. It is true that we can find out the linkage with Fold Change when we carefully see the T-Test modulation. But essentially T-Test (which represents Fold Change) is the difference of average significance between analyzed groups (this corresponds to molecule of modulation) divided by the variance among the groups. Therefore, absolute value of T-Score become bigger, when the difference among each group is smaller, and also the difference of average significance between two groups are bigger.

Bigger the absolute value of T-Score, the statistical significant will be more guaranteed.

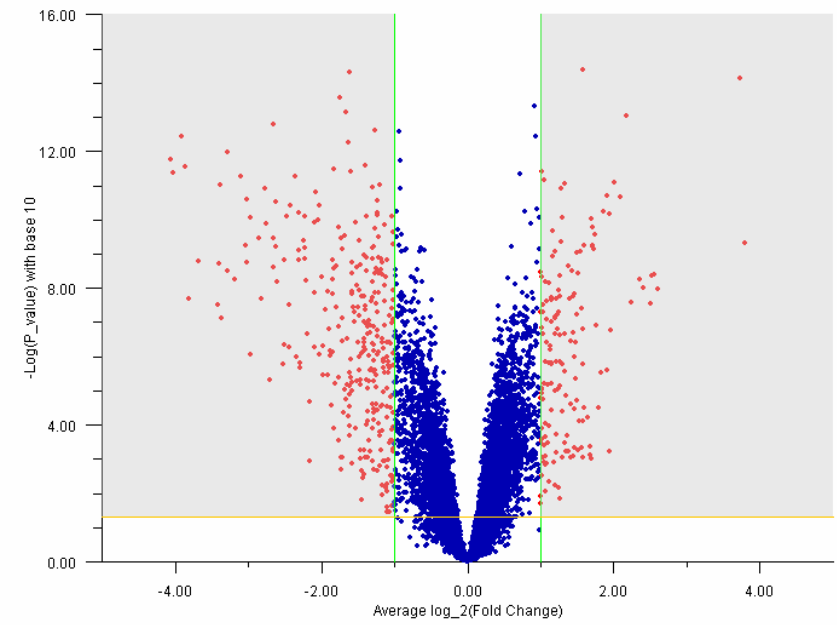
This statistical significance is known through P-Value, we can divide in ways of obtaining P-Value following the assumption of the data. The Welch approximation method is used in case of assuming that the data is following regular distribution, and generally permutation test is used when no other ratio distribution is assumed. But, we have to keep in mind that generally, to obtain the best result of T-Test, it needs to apply at least 5-6 or more replications among each group.

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \quad \rightarrow \text{approximately t - distributed with d.o.f, } \upsilon$$

$$\text{where } \upsilon = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1 - 1) + (S_2^2/n_2)^2/(n_2 - 1)}$$

7.3. Volcano Plot

This name is given because it looks like the eruption of volcano. This is a useful visualization method to see the distribution in one view, the genes extracted in Fold Change method and T-Test method. For example, among more than 2-fold DEG, statistically expressed genes (small P-Value) are our concern. To select these genes, we will have to be concerned on genes in the corner of the upper side of the figure as seen below (grey part of the figure below).



7.4. Analysis of Variance (ANOVA)

The experimental design for DEG finding, it does not have to have always 2 groups to compare. For example, if there are 2 groups to compare, there is no problem to apply Fold Change or T-Test method, but if there are more than 3 groups, what shall we do? There are 2 ways to solve this problem.

First, apply the T-Test to all possible pairs, second, apply ANOVA to all groups in one time.

For example, if there are 7 groups to compare, there will be 21 pairs to analyze when applying the first way, and numerous DEG lists will out come from each pair.

But if it is to find out DEG which shows significantly different in statistical meaning among 7 groups, this method is not the appropriate way.

Even if the statistical significance level is set up in p=0.05 for each 21 T-Test, it is possible to expect to be false positive for approximately $21*0.05 \cong 1$ Test result.

Accordingly, in case there are more than 3 groups to compare, the statistical significance level that the user has set up is guaranteed, and the useful way of analyzing at once is ANOVA method

8. Clustering Algorithm

This is the method to clustering genes or sample following similar significant pattern, the former one is called gene clustering and the other one is called sample clustering.

Gene clustering is used for gene function search, and sample clustering is used for diagnosis, prognosis and prediction of disease in clinical field

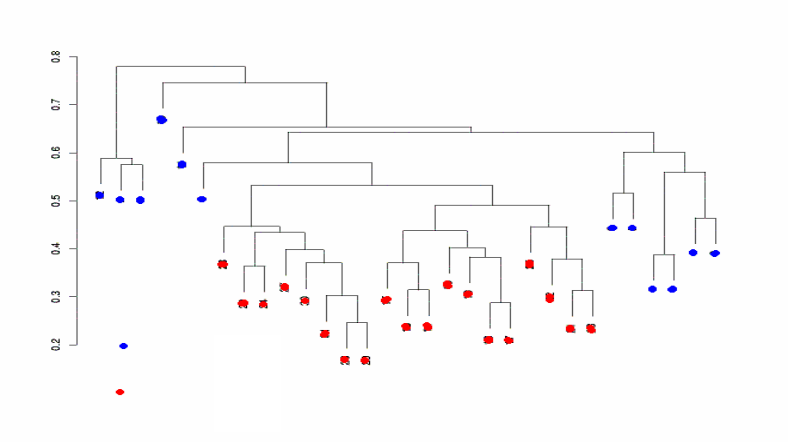
8.1. Hierarchical Clustering (HC)

Hierarchical Clustering is a classical and a general Clustering Algorithm used in statistics. This gene clustering method which used broadly after Eisen et. al thesis that is a study of external stimulus of yeast molecule genetic reaction through DNA Chip.

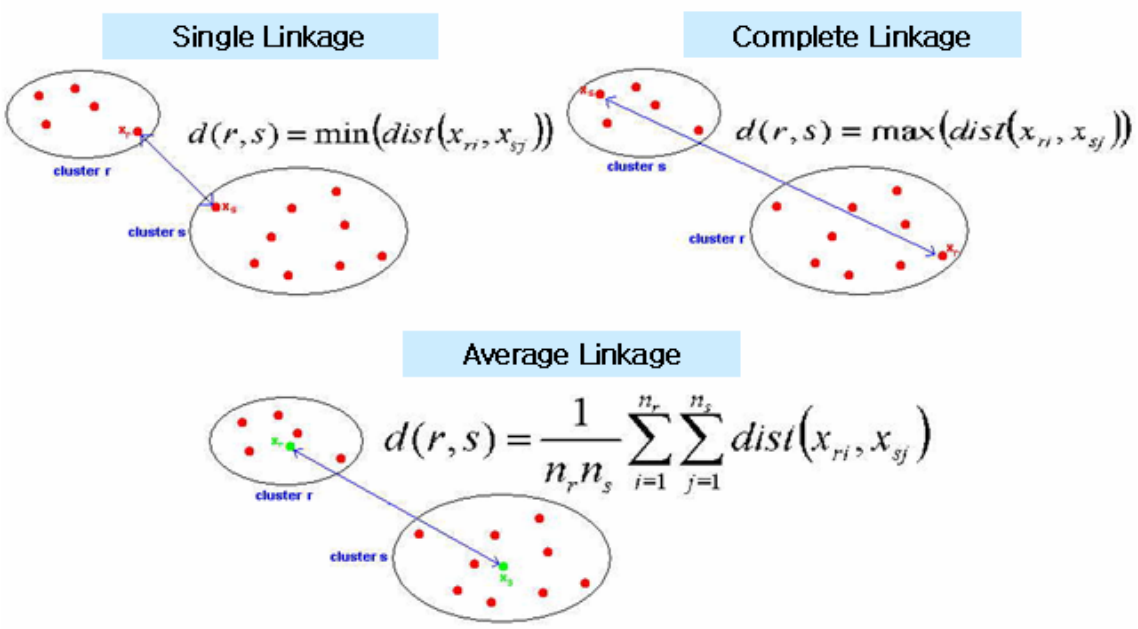
Hierarchical Clustering can be divided in Divisive Approach and Agglomerative Approach, but Agglomerative Approach is generally used. Divisive Approach is called top-down method because it approaches from the bigger group to detailed group, and Agglomerative Approach is called bottom-up method because it approaches cluster from nearest individuals to the bigger group.

Followings are gene clustering method using Hierarchical Clustering. For example, suppose there are 1,000 genes.

- First Step: Algorithm activates considering each gene in one cluster.
- Second Step: Cluster in one, after finding most similar two clusters in significant pattern among 1,000 clusters. This procedure leaves us 999 clusters.
- Recalculate the similarity value and cluster in one, after finding most similar two clusters in significant pattern among 999 clusters. This procedure leaves us 998 clusters. Repeat this procedure to 999^{th} step, finally one cluster will be left. And the result of this clustering will be shown in Dendrogram of a tree format (figure below).



One thing we have to notice from the above Algorithm. That is, how much is it near between two clusters? In other words, how define the similarity and the dissimilarity. Following this definition, linkage type and distance measure of two clusters will be fixed. Among the linkage method, Single Linkage method is a procedure to renovate with entire cluster similarity selecting high similarity value with the cluster of counterpart among former clusters composing new cluster. Complete Linkage method is a procedure to renovate with entire cluster similarity selecting low similarity value with the cluster of counterpart among former clusters composing new cluster. Average Linkage method is a procedure to renovate with entire cluster similarity calculating the average similarity with two former clusters each and counterpart cluster composing new cluster.



As Distance Measure, there are Euclidean, Minkowski, Mahalanobis Distance, and they can be shown as following formula. The distance to compare for gene i and j is shown as d_{ij}.

Let's say X for gene information, which the number of gene is p, number of sample is n, and define the distance between two optional genes as;

$X_i^R = (x_{i1}, x_{i2}, ..., x_{in})$과 $X_j^R = (x_{j1}, x_{j2}, ..., x_{jn})$ 사이의 거리를 d_{ij}^R이라고 정의하자.

■ Euclidean Distance

$$d_{ij}^R = \sqrt{(X_i^R - X_j^R)^T (X_i^R - X_j^R)} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Euclidean Distance shows actual distance used most generally.

■ Minkowski Distance

$$d_{ij}^R = \left[\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^m \right]^{\frac{1}{m}}$$

This Distance is the distance considering dimension information belonging to the individual.

■ Mahalanobis Distance

$$d_{ij}^R = \sqrt{(X_i^R - X_j^R)^T S^{-1} (X_i^R - X_j^R)}$$

This Distance is the statistical distance between two genes. This becomes Euclidean Distance when Identical Matrix is S.

■ Correlation Coefficient

$$\rho_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_{i.})(x_{jk} - \bar{x}_{j.})}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_{i.})^2} \sqrt{\sum_{k=1}^{n} (x_{jk} - \bar{x}_{j.})^2}}$$

The strong point of Hierarchical Clustering is to show in visualization, and when it is clustering, it is no need to input directly the parameter value. That is to say, there is no need to input the estimated number of cluster in advance like K-means or SOM. Also, in Dendrogram, there is a good point that we can fix the size and number of cluster that user desires. In other side, the weak point of Hierarchical Clustering is, when once clustered in each step, in further step, because of remains without going through refinement procedure, the tightness of each cluster can be less than K-means method. So clustering result cannot be satisfied than other method

8.2. K-means

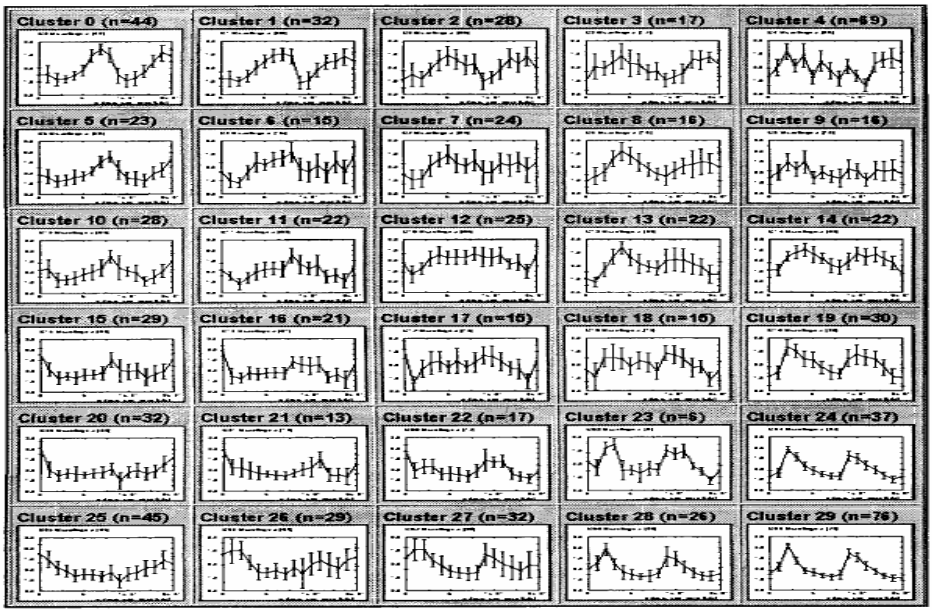
This is the method to find out the cluster of the optimum K unit through repeating calculation procedure. It operates the repeating procedure till it reaches to a certain level based on the judgment how much the constituent (it means gene in case of gene clustering) of each cluster is massed in each central group (centroid: it means average vector mathematically).

The strong point of K-means is that the resulting clusters are relatively good in clustering together in operating mathematic optimizing through repeating procedure. But, the user have to input the unit (K) of cluster in advance, and the result can come out differently following the given method of centroid of K unit given initially

8.3. Self Organizing Map (SOM)

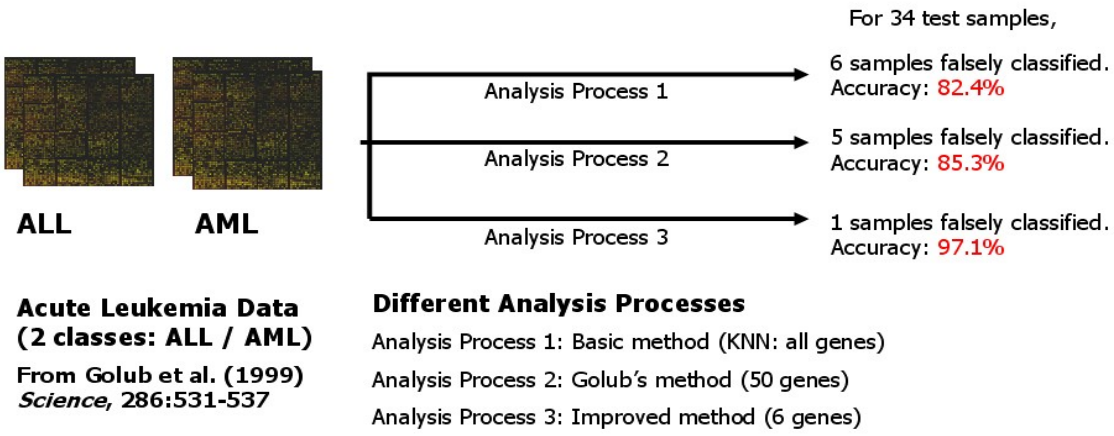
SOM is the method relatively developed recently in Computer Science field and is used broadly in other fields. This is used generally after publication of Tamayo et al., and Golub et al. of DNA Chip analysis. The most strong point of SOM is that, we can consider this, as a high-dimension data transformed in to low-dimension (generally 2 Dimension) to see it visually. This characteristic has given help in analyzing high-dimension DNA Chip data. Also, SOM can be understood as the generalized format of K-means, the user can control the parameter value, and can have desired result format. But, this point can rather be annoying to biologists.

The good point of SOM which is the visualization, is to show the similar cluster pattern in neighborhood



9. Classification Algorithm

There are some people who think that if only DNA Chip experiment is successful, the result can be easily translated with basic procedure without any effort. Let's say, they think if the material is the best, the food will be tasty. But, even if you have best material, it has to go through best cook's hands, and then the food will taste delicious with best flavor of the material. In case of DNA Chip is the same. It needs to go through detailed analyzer's hands. For a good example, there is a method of sample Classification analysis which is called generally Classification. The figure below shows us clearly how different the result can be, following the data analysis



The ultimate goal of the Classification is to have more accurate classified result with less number of genes. To do this, it is needed to select genes (gene selection procedure) to classify that show characteristics of cluster, classify samples (classifier selection procedure), and then for the last procedure, Error Estimation for confidence (Generalization Error Estimation procedure) which is most important

9.1. Gene Selection

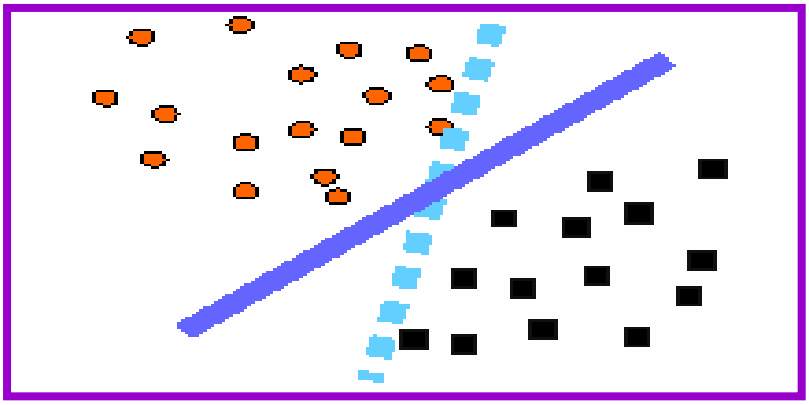
Gene Selection is a method to find out the genes which distinguish each cluster and also shows each cluster characteristically. Generally thousands and millions of gene expression figures are given to the DNA Chip. Among these genes it is the object of this procedure to find out tens and hundreds, or even several marker genes.

Gene selection method can be divided in to two. One is Uni-Variate Approach and the other one, Multi-Variate Approach. The first one is the method to select the genes with highest expression capacity after calculating individually the expression capability of each individual gene. And the second one is the method to select several genes in one time considering correlation between genes. According to the short time of calculation and expectation of effective classification result Uni-Variate Approach is generally used in gene selection, but also the Multi-Variate Approach is adopted to complement the correlation between genes which is not considered in the former method. In Multi-Variate Approach, the dimension decrease methods like PCA or SVD are used generally

9.2. Classifier

If the gene selection is done, this will be basic to classify samples. Classifying the sample this way is the Classifier, there are various methods from Fisher's Linear Discriminant Analysis (FLDA) which is used in general traditionally, to Support Vector Machine (SVM) which is the most recent way, and artificial neural network.

Let's try to understand Classifier through the figure below. Red circles are the samples of cluster 1, and black squares are the sample cluster 2. Now, let's draw a line of boundary between two groups. Following this boundary, samples in the future will be classified. Then, how do we know that we have drawn the boundary properly to divide the field of two groups? Among dotted line and solid line, which boundary is more convenient to classify cluster 1 and 2? Classifier is the answer to these kinds of questions.



9.3. Generalization Error Estimation

This is not the part which actually operates the Classifier, but this can be the most important part in Classification analysis for Error Estimation judgment standard. The core of Classification analysis is to obtain the accuracy with the classified genes and classifier. Especially, because of the classification analysis practical field is the medical field like diagnosis, prognosis and prediction, so calculation Error Estimation is very important. When we inspect the methods reported as high Error Estimation of DNA Chip data in certain thesis generally, with other similar characteristically individual data, there are few cases that show the lower Error Estimation than reported figure. In case of DNA Chip, because there are only a few numbers of samples, it is not easy to obtain the reliable Error Estimation with these samples. Thus, adequate Error Estimation method is required in this circumstance. The graph below shows that if the adequate Error Estimation method is not applied the accuracy can be pumped up.

Ambroise et. al. has mentioned the difference between external validation and internal validation of Leave-One-Out Cross Validation (LOOCV) which is generally applied, and to complement this, compared the method like Bootstrap, 10-fold CV (see the graph below).

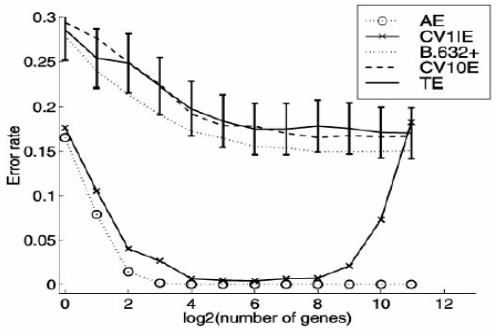


Fig. 1. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 62 colon tissue samples into training and test subsets of 31 samples each. TE, test error.

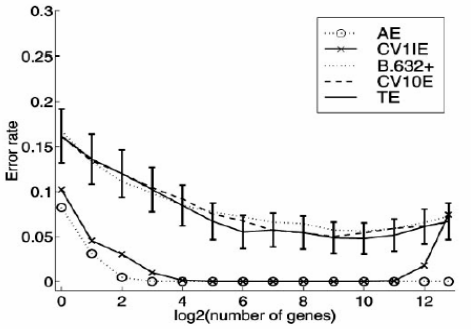


Fig. 2. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 72 leukemia tissue samples into training and test subsets of 38 and 34 samples, respectively. TE, test error.